

# Easy Come, Easy Go? Examining the Perceptions and Learning Effects of LLM-based Chatbot in the Context of Search-as-Learning

YEONSUN YANG, Electrical Engineering and Computer Science, DGIST, Republic of Korea

AHYEON SHIN, Electrical Engineering and Computer Science, DGIST, Republic of Korea

MINCHEOL KANG, Electrical Engineering and Computer Science, DGIST, Republic of Korea

JIHEON KANG, Electrical Engineering and Computer Science, DGIST, Republic of Korea

XU WANG, Computer Science and Engineering, University of Michigan, USA

JEAN Y. SONG, Humanities, Arts, and Social Sciences, Yonsei University, Republic of Korea

The cognitive process of Search-as-Learning (SAL) is most effective when searching promotes active encoding of information. The rise of LLMs-based chatbots, which provide instant answers, introduces a trade-off between efficiency and depth of processing. Such answer-centric approaches accelerate information access, but they also raise concerns about shallower learning. To examine these issues in the context of SAL, we conducted a large-scale survey of educators and students to capture perceived risks and benefits of LLM-based chatbots. In addition, we adopted the encoding-storage paradigm to design a within-subjects experiment, where participants (N=92) engaged in SAL tasks using three different modalities: books, search engines, and chatbots. Our findings provide a counterintuitive insight into stakeholder concerns: while LLM-based chatbots and search engines validated perceived benefits on learning efficiency by outperforming book-based search in immediate conceptual understanding, they did not result in a long-term inferiority as feared. Our study provides insights for designing human-AI collaborative learning systems that promote cognitive engagement by balancing learning efficiency and long-term knowledge retention.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**.

Additional Key Words and Phrases: Search-as-Learning, Human-AI Interaction, Education and AI, LLM agents

## 1 INTRODUCTION

Search-as-Learning (SAL) [65] represents a powerful method for knowledge acquisition when harnessed effectively. During an active SAL process, learners engage in the natural flow of information processing [45], as they move through the critical stages of encoding, storage, and retrieval to build a complete mental model of a concept [25, 26, 71]. Traditionally, this process has relied on conventional resources such as web search engines, academic databases, libraries, and static materials like textbooks [34, 39, 82]. These established modalities often require significant cognitive effort, as learners must actively structure and synthesize information on their own [78]. For example, a learner using a textbook for SAL must actively navigate the table of contents and index, cross-reference concepts, and synthesize information from different chapters to form a coherent understanding. Similarly, when using web search engines, the learner must evaluate the credibility of multiple sources, compare conflicting information, and organize disparate pieces of text, images, and videos.

In contrast to these traditional methods, recent advances in Large Language Models (LLMs) and conversational chatbots introduce a fundamentally different learning experience [36, 60]. Rather than merely retrieving lists of

---

Authors' Contact Information: [Yeonsun Yang](#), Electrical Engineering and Computer Science, DGIST, Daegu, Republic of Korea, [diddustjs98@dgist.ac.kr](mailto:diddustjs98@dgist.ac.kr); [Ahyeon Shin](#), Electrical Engineering and Computer Science, DGIST, Daegu, Republic of Korea, [ahyeon@dgist.ac.kr](mailto:ahyeon@dgist.ac.kr); [Mincheol Kang](#), Electrical Engineering and Computer Science, DGIST, Daegu, Republic of Korea, [presidentmc@dgist.ac.kr](mailto:presidentmc@dgist.ac.kr); [Jiheon Kang](#), Electrical Engineering and Computer Science, DGIST, Daegu, Republic of Korea, [kangjiheon@dgist.ac.kr](mailto:kangjiheon@dgist.ac.kr); [Xu Wang](#), Computer Science and Engineering, University of Michigan, Ann Arbor, Michigan, USA, [xwanghci@umich.edu](mailto:xwanghci@umich.edu); [Jean Y. Song](#), Humanities, Arts, and Social Sciences, Yonsei University, Incheon, Republic of Korea, [jeansong@yonsei.ac.kr](mailto:jeansong@yonsei.ac.kr).

information based on keywords, these tools act as collaborative assistants [49], delivering synthesized summaries and engaging in dialogue. The shift to conversational chatbots makes the search easier and more “answer-centric,” introducing a fundamental trade-off: higher efficiency in information access versus the potential for shallower knowledge gain. Although learners can achieve higher information throughput [63] by acquiring more facts per unit of time, this efficiency may come at the expense of long-term learning retention [35] and the formation of robust knowledge structures [53].

Despite the promise [58, 61] and accompanying concerns [41, 43] about the use of LLM-based tools in education, there remains a lack of empirical research that systematically evaluates their effectiveness in the context of SAL. We believe that this gap exists because the roles of LLM-based chatbots—including how they are perceived by both instructors and learners and their actual impact on learning outcomes—have not been adequately compared to traditional SAL tools like textbooks and search engines. To address this, our research focuses on a systematic investigation of LLM-based chatbots as a new tool for SAL. By specifically comparing their effects with those of traditional tools, we aim to uncover the unique contributions and drawbacks of chatbots and investigate how their role might differ from that of conventional methods. We expect that this comparative approach can offer crucial insight into how LLMs can best support SAL practices and inform the design and development of effective learning strategies for human-AI collaborative mechanisms.

To this end, we conducted a large-scale investigation comparing three SAL modalities: textbooks, a web search engine (Google [31]), and an LLM-based chatbot (ChatGPT [55]) during July 2024. Through large-scale surveys with educators ( $N = 75$ ) and university students ( $N = 92$ ), we first provide a comparative overview of the *perceived* potentials and pitfalls of using LLMs for learning. We then complement these findings with a within-subjects experiment inspired by the encoding-storage paradigm [37], where the 92 participants engaged in SAL tasks using each of the three modalities in a randomized order. The tasks involved a sequence of activities including information interaction, reviewing, concept map drawing, and closed-book testing. We examined how the three modalities affect learning outcomes in terms of learning efficiency, information throughput and accuracy, and long-term knowledge retention. The results reveal a trade-off between immediate learning gain and longer-term retention: while the chatbot condition facilitated the highest immediate recall of concepts and inter-conceptual links, and together with the search engine, achieved significantly higher immediate test scores than the book condition, these benefits were rather temporary. The lack of significant differences on the two-week delayed closed-book tests of the three modalities indicates that this initial performance boost did not translate into superior long-term retention. Further analysis based on Bloom’s taxonomy show that the immediate performance benefits of the LLM-based chatbot and search engine were confined to the *Understand* level, with no significant differences observed in the *Apply* or *Analyze* levels compared to text books.

Interestingly, our findings challenge the validity of educators’ concerns regarding limited cognitive engagement in LLM-based chatbots for SAL. While survey respondents worried that the passive nature of chatbot interaction would compromise the depth of learning and internalization, the comparable long-term retention scores across chatbot, search engine, and book conditions indicate otherwise. The anticipated *negative impact* on knowledge retention was not observed; rather, retention rates were similar across all conditions, even as the chatbot and search engine provided the efficient and adaptive support users expected. Although the initial superiority of digital tools was transient, it did not lead to a knowledge deficit in the long run compared to traditional text book reading. This suggests that while AI and search engines facilitate rapid information intake through lower cognitive load, they do not appear to compromise the ultimate volume of retained knowledge. We discuss whether this “easy come, easy go” phenomenon stems from cognitive retention ceilings or the absence of desirable difficulty, highlighting the need for future designs that balance the ease of AI assistance with the cognitive engagement required for consolidation.

## 2 RELATED WORK

### 2.1 Information Processing Theory and SAL

The theoretical foundation for our work is rooted in information processing theory [25, 71], a cognitive framework that models the human mind as a system for processing, storing, and retrieving information. Information processing theory posits that learning is not a passive event but an active process where learners move information through distinct memory stages: sensory memory, working memory, and long-term memory [3]. This perspective emphasizes that the depth of learning is directly tied to the level of cognitive effort, or mental investment, required to move and encode information from short-term to long-term memory [56].

This view of learning is particularly relevant to SAL, which has long been a foundational concept in cognitive science, examining how learners transform information-seeking into knowledge acquisition [46, 72, 78]. Historically, this framework was primarily built around traditional tools that provide static information. Early SAL studies explored how learners navigate and synthesize content from compiled resources like textbooks and encyclopedias [22, 67, 69]. The advent of web search engines (e.g., Google [31], Bing [47]) marked a significant shift, offering vast but disparate information that required learners to shoulder a substantial cognitive burden for source evaluation and synthesis [24, 75, 79]. Because the act of navigating, evaluating, and integrating disparate information pieces is what gives the SAL process its learning value, a traditional SAL experience aligns with the principles of information processing theory. The inherent difficulty of the task forces the learner to exert the cognitive effort necessary for higher-order thinking.

However, the recent emergence of conversational chatbots based on LLMs represents a fundamentally new inflection point, prompting us to reconsider how to best support SAL in this new era of AI. Unlike traditional tools, chatbots serve as collaborative assistants [4, 18, 36], delivering synthesized information and engaging in dialogue. This shifts the learner's role from a passive retriever to an active collaborator, presenting a new set of challenges that extend beyond simple information access. Our study directly addresses this need by empirically comparing the effectiveness of LLM-based chatbots against traditional tools: textbooks and search engines. Through this comparative approach, our goal is to identify the unique benefits and drawbacks of the modalities and ultimately propose design guidelines for AI-mediated learning environments that cultivate effective SAL practices.

### 2.2 LLMs in Education: A Dual Perspective on Promise and Peril

Recent advances in artificial intelligence (AI), particularly LLMs, have led to their increasing use in cognitively challenging tasks. These tools have the potential to significantly enhance productivity in many domains, including writing [18] and other collaborative tasks [30, 70]. However, applying this technology to education raises considerable concerns [41, 43], as the core goal of learning is not just to produce correct answers but to develop critical thinking skills and deeper knowledge. This pursuit of intellectual growth requires significant cognitive engagement, a process that an over-reliance on AI assistance could compromise. In this paper, we investigate this tension in the context of SAL, a primary mode of self-directed knowledge acquisition.

Prior research on the use of LLMs in education has presented a dual narrative of promise and peril. On one hand, studies have demonstrated the potential of LLMs to act as personalized tutors [13, 57], generate customized learning content [20], and provide instant feedback [62], leading to increased efficiency and motivation [44]. These applications suggest that LLMs could streamline the learning process and make it more accessible. On the other hand, a growing body of work has raised critical concerns. They highlight risks such as the potential for LLMs to generate inaccurate or hallucinated information, which can mislead learners and undermine the quality of knowledge acquisition [5, 58].

Additionally, researchers have voiced concerns that over-reliance on LLMs may impede the development of critical thinking, problem-solving, and information literacy skills that are central to higher-order thinking [21, 50, 61, 81]. While these studies provide a valuable conceptual foundation for understanding the opportunities and risks, they often lack a direct, empirical comparison of how LLMs’ unique capabilities and features translate into concrete learning outcomes compared to different SAL tools.

Our study directly addresses these research gaps by employing a large-scale within-subject design that compares three distinct SAL tools, integrating both student and educator perceptions with empirical metrics of efficiency, accuracy, and long-term retention to offer a comprehensive understanding of LLM-based chatbots for SAL.

### 3 STUDY

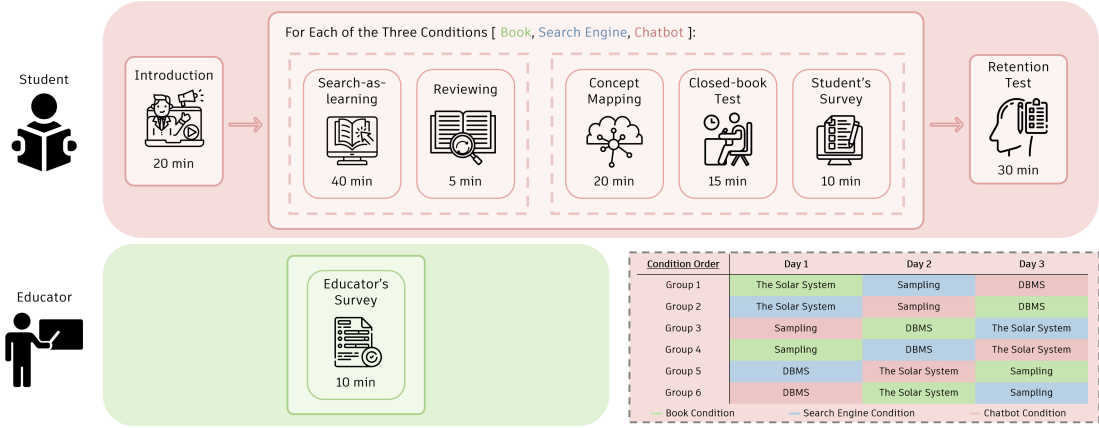


Fig. 1. Procedure over the five steps in a mixed-method within-subject study with 92 university students.

To systematically investigate the effectiveness of LLM-based chatbots for SAL, we adopt a mixed-methods approach that structures the exploration of both perceived and empirical effects (Figure 3). Specifically, we surveyed educators ( $N = 75$ ; E1–E75) and students ( $N = 92$ ; S1–S92) to gain a comprehensive understanding of both pedagogical (teaching) and practical (learning) perspectives. In addition, we conducted a controlled within-subject user study with students ( $N = 92$ ) to provide empirical evidence on learning outcomes, focusing on search efficiency, accuracy of acquired knowledge, and delayed retention. To draw complementary insights, we conducted the student survey after the user study, where they were asked to reflect on their learning experience with each condition. The educator survey was conducted concurrently with the user study to gather their perspectives. Note that the surveys and experiment were conducted during July 2024. The study was approved by the Institutional Review Board at the author’s university.

Our research questions are as follows:

- RQ1: [Perception] What are the main benefits and risks perceived by students and educators when using LLM-based chatbots for SAL?
- RQ2: [Efficiency] How do the completion time and the number of search Q&A pairs differ between using LLM-based chatbots and traditional search methods?
- RQ3: [Accuracy] How does the information throughput and accuracy of knowledge acquired with LLM-based chatbots compare to that acquired with traditional search methods?

- RQ4: [Retention] How does the long-term retention of knowledge acquired with LLM-based chatbots compare to that acquired with traditional search methods?

### 3.1 Survey on Perceptions of SAL Tools: Textbooks, Web Search Engines, and LLM-based Chatbots

**Survey with Educators.** We recruited 75 South Korean educators with varying educational backgrounds and experiences through snowball sampling and online advertising. All educators work in educational institutes in South Korea, except for two who are based in the United States. Our respondents included 33% females, 64% males, and 3% who did not disclose their gender. The age distribution was as follows: 25% were 20–29 years old, 24% were 30–39 years old, 29% were 40–49 years old, and 21% were 50 years or older. Our sample consisted mainly of university or high school teachers (93% of the sample). Most of them either work in the STEM field (Science, Technology, Engineering, Mathematics; 57%) or in the HSS field (Humanities and Social Sciences; 33%) and 63% of them had experience using LLM-based chatbots. The survey took about 10 minutes to complete. Educator participation was voluntary, with no monetary compensation.

In the survey, open-ended questions were asked to explore their perspectives on the benefits and drawbacks of LLM-based chatbots in the SAL process, *e.g.*, their thoughts on the benefits and risks of using LLM-based chatbots. They were further asked to rank the three tools in terms of which they would encourage students to use when studying unfamiliar concepts independently in courses, along with their rationale. We used a rank instead of a Likert scale to make direct comparisons between tools, revealing their relative preferences and priorities, and to avoid the tendency to give neutral answers [33, 83].

**Survey with Students.** In a post-survey after the experiment, students were asked to reflect on their SAL experiences with books, search engines, and an LLM-based chatbots. The survey included open-ended questions for comparative evaluations of the three tools, their intended future use in university coursework, and the perceived benefits and risks of each tool with supporting rationales. The survey took about 10 minutes to complete. Compensation for the survey was included in the total payment for the entire experiment.

### 3.2 Experiment Design

We performed a within-subject experiment with 92 university students, where they engaged in self-directed learning to achieve predefined learning objectives in the three SAL conditions. The study was conducted in a university-level STEM context, given that STEM texts typically involve complex structures and technical terminology that place high demands on learners’ cognitive resources. To allow direct comparison across conditions, we designed the study in a web-based environment. As the experimental apparatus, we developed a browser extension to log search throughput (*i.e.*, the number of search Q&A pairs per unit time). We introduced learning tasks to assess immediate outcomes and delayed retention in learning. To minimize fatigue and order effects, we used a counterbalanced design in which the three modalities and three STEM modules were rotated across participants, resulting in six sequence groups (Figure 1).

Table 1. Comparison of SAL conditions across four dimensions.

Condition	Information Throughput	Learner’s structuring and encoding burden	Information quality	External storage quality
Book	Low	High	High (reliable, structured)	Low–Medium (self-made)
Search Engine	Medium	Medium	Low–High (uncertified)	Medium
Chatbot	High	Low	Low–High (hallucination)	High (pre-synthesized)

**3.2.1 Search-as-Learning Conditions.** Inspired by the encoding-storage paradigm [14, 37], we designed three experimental conditions (Table 1). The **Book** condition replicates academic textbooks and scholarly publications, which is characterized by its **low information throughput**. Accessing specific information within a physical book requires a sequential and manual process of locating the text, navigating chapters, and scanning pages—notably slower than digital alternatives. Despite this, textbooks impose a **high structuring and encoding burden for learners** [40, 74]. While the content itself is typically highly structured and logically presented by experts, learners must actively engage in deep reading, comprehension, and critical analysis to internalize, summarize, and integrate this knowledge into their own cognitive frameworks. This active processing is crucial for effective learning, but demands significant cognitive effort. Correspondingly, textbooks consistently have **high information quality**, resulting from rigorous editorial, peer-review, and fact-checking processes that ensure accuracy, reliability, and conceptual coherence. Finally, textbooks typically lead to **low to medium external storage quality**. Although the physical book serves as a robust external repository, its utility as personal external storage (e.g. for quick recall or reference) is largely dependent on the learner’s own efforts to create supplementary aids such as notes, highlights, or summaries. The effectiveness of this self-constructed external storage varies significantly with the learner’s skill and diligence.

The **Search Engine** condition offer a **medium information throughput**, providing rapid access to a vast index of online resources through keywords and query based searches. Queries yield numerous links almost instantaneously, allowing for quick initial retrieval. However, the learner must still manually navigate through various webpages, filter irrelevant results, and critically evaluate the information, which slows down the overall processing compared to direct answers. This process contributes to a **medium structuring and encoding burden for learners**. Learners are responsible for sifting through fragmented and often conflicting information from diverse sources, assessing credibility, and then synthesizing these disparate pieces into a coherent understanding. Search engine results can vary in **low to high information quality** because most content are uncertified, where anyone can create content online. This wide range reflects the open nature of the internet. Content can vary from highly reliable, peer-reviewed articles to unsubstantiated claims, or misinformation. The absence of a universal vetting mechanism for all indexed content means that learners must exercise significant discretion and critical evaluation. Consequently, the **external storage quality is medium**. While learners can leverage browser functionalities (e.g. bookmarks, tab categorization) or dedicated tools to manage searched pages, the engine itself provides uncured and disparate information. This requires that the learners personally organize and synthesize this content into genuinely usable and coherent external knowledge.

The **Chatbot** condition provides **high information throughput**, delivering direct, synthesized *answers* and summaries to complex queries almost instantly. This eliminates the need for users to navigate multiple sources or manually filter results, thus maximizing the rate at which relevant information is presented. Consequently, the **learner’s structuring and encoding burden is low**. Chatbots provide information that is largely pre-digested, often structured into coherent paragraphs or lists, which significantly reduces the cognitive effort required for the learner to structure and integrate new knowledge internally. However, the **information quality spans low to high** mainly due to the hallucination effect. While they can synthesize information into accurate and insightful responses (high quality), they are prone to generating confident yet factually incorrect or entirely fabricated information (low quality). This inherent unpredictability requires careful verification by the user. Finally, the **external storage quality is high** because chatbots provide information that is already processed, summarized, and often structured in a ready-to-use format. This pre-synthesized knowledge minimizes the learner’s effort in finding, organizing, and preparing information, effectively serving as an immediate and accessible extension of their own cognitive resources.

For the experimental setup, we selected a predefined textbook [12, 38, 59] for **Book** condition, Google [31] for **Search Engine** condition, and ChatGPT-4o [55] for **Chatbot** condition as representative tools.

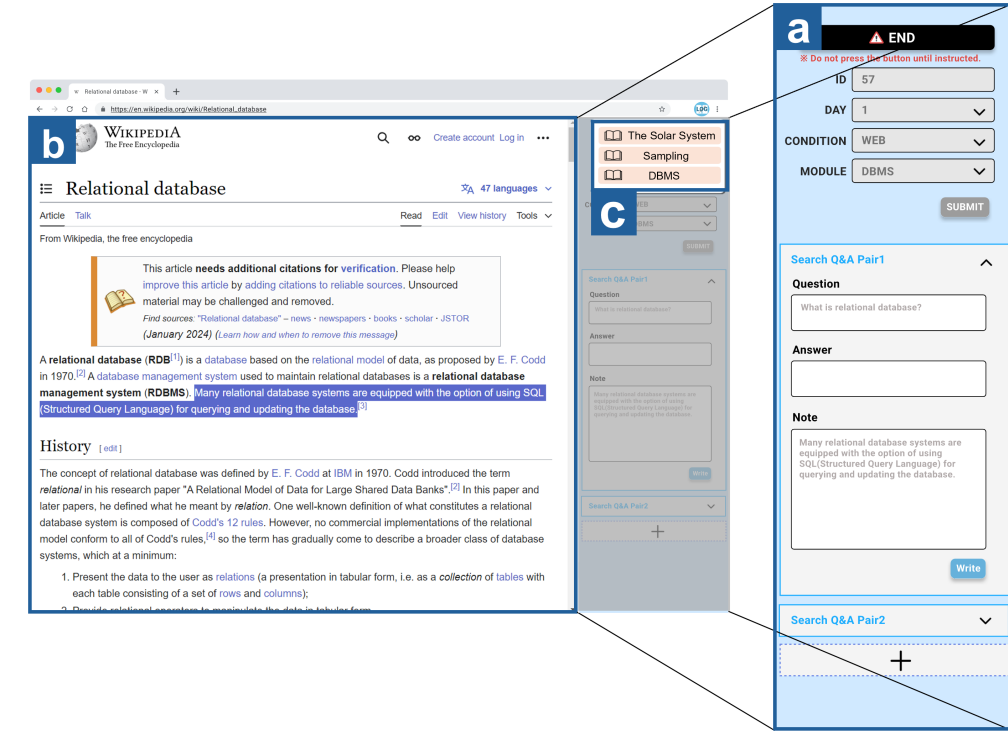


Fig. 2. Overview of the SAL logger of **Search Engine**. The interfaces for **Book** and **Chatbot** follow a similar structure, consisting of two interface components: **a**, the content panel and **b**, the note panel. A Chrome extension plug-in **c** was provided for the **Book** condition to access pre-selected textbooks in PDF format. The experiment starts by inputting information such as participant ID, session (from Day 1 to Day 3), experiment condition, and study subject on top of the note panel. Participants start by initiating a question themselves in the Question section. As they engage in self-guided search and learning on the content panel, they can drag relevant information or organize it into their own words in the Note section. Once they find an answer to their query, participants complete the Answer section and submit it. They can add new notes by clicking the '+' button.

**3.2.2 Experimental Apparatus.** The SAL Logger (Figure 2) runs on Chrome browser and records timestamps of user interactions, including participants' queries and accessed resources. We designed the logger to switch among the three conditions of SAL tools introduced in Section 3.2.1 based on an experiment condition variable. The logger consisted of two interface components: a content panel (Figure 2-a), where participants interact with the assigned SAL tool to perform tasks, and a note panel on the right (Figure 2-b), where they collect and organize information within each SAL search Q&A pair. We developed the SAL logger as a Chrome extension using HTML and JavaScript. It has a client-server architecture that stores search histories, and collects learning logs. A dedicated server was implemented to ensure the security of learning data and fulfill requirements for robust client-server operation within our experimental framework. The server is implemented using Node.js and stores data in JSON format.

Our system supports the three experimental conditions as follows:



- **Book**: The textbooks were sourced from university-level courses and provided in PDF format, covering all predefined learning objectives (The Solar System [12], 568 pages; Sampling [38], 461 pages; DBMS [59], 228 pages). They were translated into Korean to remove language barriers and integrated into our SAL Logger (Figure 2-c). To distinguish this condition from web documents, participants were instructed to use the table of contents and index instead of full-text search, reflecting the high encoding effort of book-based search.
- **Search Engine**: The browser window displays multiple web documents, blogs, or other resources. We restricted access to Google [31] to provide a consistent search experience, limited to traditional search results without LLM integration (as of July 2024). To control for personalization due to search history, all searches were conducted using a dedicated experimental account.
- **Chatbot**: The content panel displayed the ChatGPT interface for conversational search. We restricted this condition to ChatGPT-4o [55], selected for its enhanced speed and multimodal capabilities. To isolate and assess the intrinsic characteristics of chatbot-only interactions, we deactivated the built-in web browsing feature.

**3.2.3 Participants and Procedure.** We recruited 92 university students from our institute located in South Korea, through mailing lists and online community advertisements (average age=21, 46 males and 46 females). To mitigate potential confounding effects of prior knowledge on study results, we selected STEM modules specifically not covered in our institute’s curriculum. Furthermore, a screening process was implemented to actively exclude participants who possessed pre-existing familiarity with the target topics. The screening test consisted of three MCQs at the Remember level [42] of Bloom’s taxonomy to assess prior knowledge, as this level requires simple recall of factual information. Students who scored full marks (3 out of 3) on the screening test were filtered out. Participants were randomly assigned to one of six possible counterbalanced sequences of conditions (shown in the bottom left box in Figure 1). The average score of the screening test of all study participants was 0.55 ( $SD = 0.68$ ), with no significant differences (Kruskal–Wallis  $H = 0.675$ ,  $p = 0.70$ ) across the condition orders, confirming the successful mitigation of prior knowledge bias across all sequences.

The experiment began with an introduction to the study, providing participants with a thorough explanation of the tasks and procedures. Participants were asked to submit informed consent and received an online link to the SAL Logger. They then installed the SAL Logger on the Chrome browser and started studying the predefined learning objectives (Table 2) using a designated SAL tool. During the 40-minute study, participants formulated their own questions in the note panel whenever they identified knowledge gaps or had internal questions. To address these questions, they searched for information using the SAL tool and organized key insights or pertinent content in the Note section. Once questions were resolved, participants submitted their search Q&A pairs with answers through the system. After completing the SAL task, they were given five minutes to review and reflect on their search Q&A pairs on the SAL Logger.

Then they completed a 20-minute concept map drawing task without access to any learning materials, hand-drawing labeled nodes and linking them to illustrate relationships among concepts. Participants were instructed to construct dense and comprehensive concept maps. This task was included to capture the effective information throughput achieved by each modality. After completing the concept map drawing task, participants conducted a closed-book test, where they had 15 minutes to solve nine MCQs. Finally, participants completed a survey via an online form, which included questions about their experiences and perceptions of the given SAL condition.



This process was repeated over **three consecutive days**, wherein both the study topics and the assigned SAL tools were systematically varied (please refer to Figure 1). Lastly, two weeks later, participants completed a 30-minute retention test. Participants received 66,000 KRW (approximately 45 USD) as compensation.

Table 2. Nine learning objectives used in the within-subject controlled experiment.

Bloom's taxonomy	The Solar System	Sampling	DBMS
Understand	Define and classify celestial bodies in the Solar System.	Define the concept of sampling.	Define the concepts of databases and tables.
Apply	Explain and compare 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.
Analyze	Analyze planetary motion using Kepler's Laws.	Classify various probability sampling techniques and apply their formulas.	Analyze data with CRUD operations using MySQL and basic syntax.

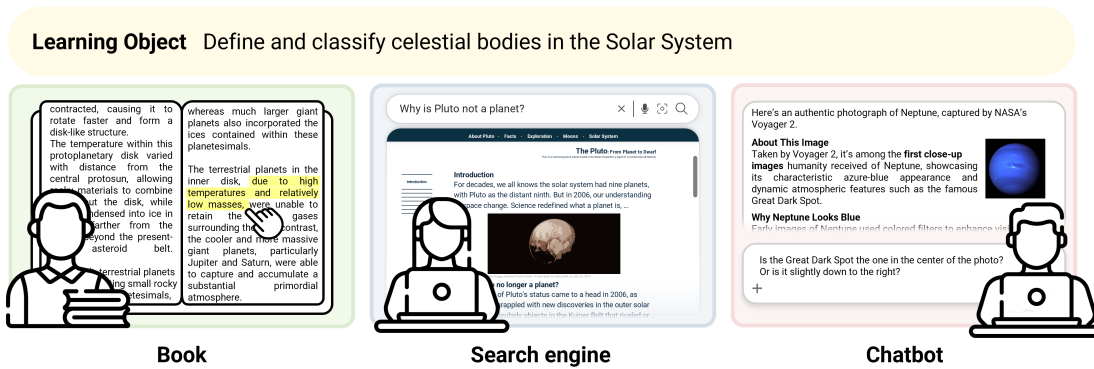


Fig. 3. Despite a shared learning objective, students' engagement patterns, interaction types, and the resultant cognitive burdens can diverge significantly across learning tools. By comparatively analyzing different SAL tools: books, search engines, and chatbots, we aim to clarify chatbots' unique contributions and limitations, offering insights into better LLM support in SAL practices. (The images are from actual experiment logs: book search highlight of S88, search engine query of S69, and chatbot log of S36.)

**3.2.4 Learning Tasks and Materials.** To control for participants' prior knowledge, we selected STEM modules not included in our institute's curriculum. The authors conducted three iterative rounds of discussion to set three modules from distinct STEM domains: The Solar System (Astronomy), Sampling (Statistics), and Database Management Systems (Computer Science). Drawing on university-level coursework available from online platforms such as Coursera<sup>1</sup> and edX<sup>2</sup>, we derived three learning objectives for each module, aligned with the Understand, Apply, and Analyze levels of Bloom's taxonomy [9, 42] (Table 2). Participants had to study all three learning objectives for each module assigned for the day. We excluded higher-level taxonomies because they are difficult to assess within the limited learning timeframe.

<sup>1</sup><https://www.coursera.org/>

<sup>2</sup><https://www.edx.org/>

*Concept map:* To capture information throughput, we employed Novakian concept mapping [11], a method commonly used to capture learners’ mental models by representing networks of connections between related concepts [17, 51, 53, 54, 64].

*Multiple-choice questions (MCQs):* The post-test consisted of nine MCQs designed to align with the predefined learning objectives. The MCQs were first generated through few-shot prompting, and the final items were selected and refined through question-level evaluation (Table 7) involving four of the authors. To ensure the validity of the MCQs, three domain experts were recruited to conduct a quiz-level evaluation using three metrics (Table 3). The expert ratings indicated acceptable quality: The Solar System scored (2,3,3), and both Sampling and DBMS each scored (3,2,3) for structure, redundancy, and usefulness. We provide details on question generation, quality evaluation, and example MCQs in Appendix A. To further assess long-term retention, we administered a delayed test two weeks later using the same questions as the post-test, with randomized order and options.

Table 3. 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.

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)

## 4 Data Analysis Measurement

### 4.1 Survey Analysis

To compare preferences for SAL tools across educators and students, we performed Friedman’s test [28]. For post-hoc analysis, we employed Conover tests [15] with Bonferroni correction to avoid potential multiple comparison problems. The mean rank metric [2] was used to calculate the average rank, ensuring robustness for non-parametric statistical tests. In addition to analyzing perceptions of LLM-based chatbots for SAL compared to conventional tools, we conducted theoretical coding [52] on open-ended responses and comments to uncover the underlying reasons behind the tool preferences. Two of the authors reviewed the responses and classified them into perceived benefits and risks of adopting LLMs in the SAL process. Conflicts were resolved through iterative discussions, and the two authors achieved inter-rater reliability of Krippendorff’s alpha of 0.89.

### 4.2 Quantitative Outcome Measures

**4.2.1 Search and Interaction Logs.** To evaluate search efficiency, we analyzed the quantity and duration of search Q&A pairs logged through SAL Logger. We note that a completed search Q&A pair is a unit in which a single question is issued, relevant information is integrated, and an answer is formulated—representing one cycle of SAL. All timestamps were logged for each event. For comparison, one-way ANOVA tests were conducted across conditions to compare the number of search Q&A pairs and the time spent per search Q&A pair.

**4.2.2 Closed-book Concept Map Drawing Task.** To assess information throughput, we analyzed students’ concept maps using count-based network metrics—the number of nodes and the number of edges—which are commonly used in prior

literature to estimate individual learners' understanding [6, 16, 27, 80]. To apply these metrics, we reviewed all 276 maps from the participants and resolved synonym conflicts (e.g., "DB" and "Database") through iterative discussions among four of the authors. We then conducted a comparative analysis of the concept map metrics using ANOVA tests.

**4.2.3 Closed-book and Retention Tests.** To assess the accurate encoding and stable storage of essential knowledge aligned with our learning objectives, we analyzed two types of tests: a closed-book test administered immediately after each condition and a retention test conducted two weeks later. Each test was scored on a 0–9 scale (because nine MCQs were provided per test), and the results were statistically analyzed using one-way ANOVA tests to compare performance across conditions. To further analyze learning outcomes by cognitive level, we conducted one-way ANOVA tests for the three Bloom levels (Understand, Apply, Analyze) across conditions. To evaluate knowledge retention, we conducted an ANCOVA on the retention-test scores, with condition as the between-subjects factor and the closed-book test score as the covariate.

## 5 FINDINGS

### 5.1 RQ1: What are the main benefits and risks perceived by students and educators when using LLM-based chatbots for SAL?

Through our survey analysis of students and educators, we examined their preferences among three SAL tools and further investigated their perceptions of LLM-based chatbots. Our findings reveal a divergence in SAL tool preferences between educators and students, with educators favoring established tools while students show a notable preference for LLM-based chatbots. This disparity stems from a fundamental pedagogical tension: students prioritize efficient, streamlined information encoding for immediate needs, whereas educators emphasize the critical role of cognitive effort and metacognition for deep, durable learning, despite both groups acknowledging the general benefits and risks of LLM-based SAL.

**5.1.1 Preferences.** As shown in Figure 4, a Friedman test revealed a statistically significant difference in rankings across the three SAL tools ( $X^2(2) = 21.62, p < .001$ ). Post-hoc analysis showed that all three SAL tools differed significantly from one another. The majority of educators preferred books (52%; 39 of 75) and search engines (36%; 27 of 75) as search-as-learning tools, while only 12% (9 of 75) chose LLM-based chatbots their favorite. This suggests that educators remain cautious about adopting LLM-based chatbots as SAL tools, showing a clear preference for more established conventional tools. Meanwhile, students exhibited an opposite preference order to educators. Notably, the difference in preference between books and LLM-based chatbots was statistically significant ( $p < .001$ ).

Overall, our comparison highlights a marked difference in tool preferences between students and educators.

**5.1.2 Perceived Benefits and Risks.** To gain a deeper understanding of the reasons behind these tool preferences, we qualitatively analyzed survey responses on perceived benefits and risks (see Table 4). While tool preferences diverged between educators and students, both groups consistently recognized most of the benefits and risks of LLM-based SAL.

In terms of benefits, both students and educators commonly acknowledged the advantages of using LLMs for SAL. These advantages collectively illustrate a key trade-off we highlight in this work—efficient information encoding—LLMs' ability to rapidly deliver pre-synthesized and structurally organized knowledge tailored to learners' goals, thereby reducing cognitive effort during the encoding phase of learning. A distinct benefit raised exclusively by students was *selective learning*, which provided focused access to the desired information they are looking for. As S5 noted, "*One of the strengths of using LLMs was that I could quickly focus on the parts I didn't know, rather than reviewing everything.*" S9

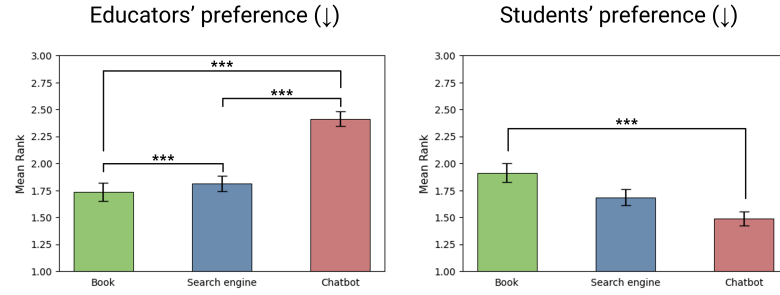


Fig. 4. Bar chart showing mean preference ranks from educator and student survey responses. A lower mean rank (↓) indicates higher preference (i.e., 1 = most preferred). The left plot (a) presents educators' rankings of the three SAL tools, and the right plot (b) shows students' rankings. Significance is based on a Friedman test and marked as  $p < .001$ \*\*\*).

Table 4. Perceived benefits and risks of LLM-based chatbots as SAL tools, reported by educators ( $N = 75$ ) and students ( $N = 92$ ). Numbers indicate how many respondents mentioned each theme.

Category	Theme	Explanation	Educators	Students
Benefits	Personalized Learning	LLMs can provide adaptive explanations and feedback tailored to learners' individual needs and progress.	26	13
	Serendipitous Discovery	LLMs can provide unexpected but relevant information beyond learners' initial queries.	8	5
	Diverse Explanations	LLMs can present the same concept through multiple explanatory styles, such as simplified summaries, detailed breakdowns, and a variety of materials.	23	29
	Time Efficiency	LLMs can provide rapid search and collect extensive information in a short time.	50	36
	Selective Learning	LLMs can provide focused access to desired information without irrelevant content.	0	14
Risks	Lack of Reliability	LLMs can provide incorrect or fabricated information without clear sources.	19	39
	Reduced Metacognition	LLMs can reduce metacognitive ability to recognize and reflect on knowledge gaps.	23	0
	Limited Cognitive Engagement	LLMs can hinder learners' direct engagement in critically evaluating, integrating, and internalizing information.	37	0
	Prior Knowledge Dependency	LLMs can provide surface-level information when learners lack sufficient prior knowledge on the topic.	0	24

also remarked, “When I searched with Google, I had to sift through tons of information to find what I needed. With ChatGPT, I could directly learn only what I was curious about, which made the process much clearer and less overwhelming.” This preference reflects a desire for streamlined and targeted learning experiences, driven by the time constraints students face in balancing academic and personal demands.

In contrast, educators valued more effortful engagement with learning materials, viewing cognitive effort as essential for deep understanding and retention. They highlighted risks such as *reduced metacognition* and *limited cognitive engagement*, warning that over-reliance on chatbots could hinder learners from critically evaluating, synthesizing, and internalizing knowledge. As E1 emphasized, “Even when students are given the same information, they develop creativity and critical thinking by making sense of it in their own way, rather than passively accepting it.”, and E8 remarked “It may

take more time, but I believe that deeply focusing on and accurately understanding core concepts is more important than quickly learning superficial facts.” However, none of the students raised this issue as a potential risk. Instead, students pointed to more immediate challenges, such as *prior knowledge dependency*, where difficulty guiding the chatbot arose without sufficient initial understanding. Nonetheless, all participants acknowledged that LLMs could be unreliable at times and emphasized the need to interpret their outputs with caution.

Taken together, these findings illustrate a pedagogical tension between short-term encoding and long-term storage. While students prioritized immediate access and ease of use, educators emphasized the importance of cognitive effort and reflective thinking in building durable knowledge.

## 5.2 RQ2: How do the completion time and the number of search Q&A pairs differ between using LLM-based chatbots and traditional search methods?

During the 40-minute search-as-learning, students completed a total of 442 search Q&A pairs with books, 502 search Q&A pairs with search engines, and 614 search Q&A pairs with LLM-based chatbots (Figure 5). On average, the number of search Q&A pairs per student was 4.59 ( $SD = 2.91$ ) in the **Book** condition (the least), 5.45 ( $SD = 3.52$ ) in the **Search Engine** condition, and 6.69 ( $SD = 5.47$ ) in the **Chatbot** condition (the most). Students completed significantly more search Q&A pairs with LLM-based chatbots than with books ( $p < .002$ ). The average time per search Q&A pair (in minutes) was longest for books ( $M = 6.87, SD = 6.57$ ), followed by search engines ( $M = 5.95, SD = 5.58$ ), and LLM-based chatbots ( $M = 4.77, SD = 4.97$ ). Students spent significantly less time per search Q&A pair in the **Chatbot** condition than in the **Book** ( $p < .001$ ) and **Search Engine** ( $p = .05$ ) conditions.

Collectively, these findings suggest that LLM-based chatbots facilitate faster and more frequent information retrieval, thereby empirically substantiating students’ perceived efficiency as SAL tools and offering a potential explanation for their strong preference for LLMs over conventional tools.

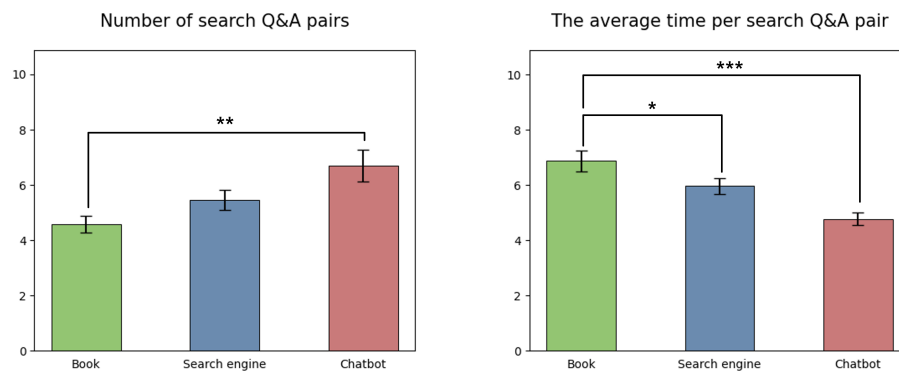
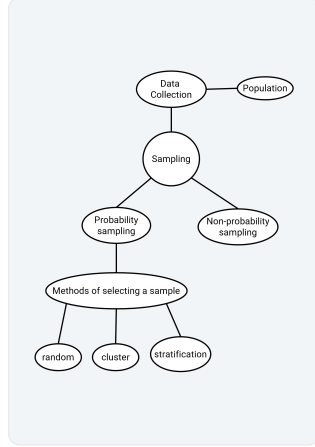


Fig. 5. Bar chart showing the number of search Q&A pairs (left) and the average time spent per Q&A pair (right). ANOVA was conducted to compare information throughput across the conditions. Significance is marked as  $p < 0.1$  (\*),  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), or  $p < 0.001$  (\*\*\*).

Example of Poor Concept Map (S88)



Example of Rich Concept Map (S10)

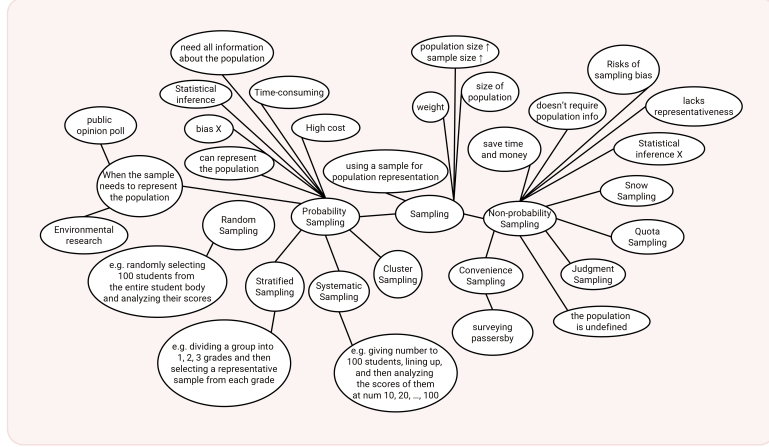


Fig. 6. Example of concept maps drawn. The left one has less information compared to the right one. Both are created for the Sampling module (left: search engine condition (S92), right: chatbot condition (S10)). This shows the version before resolving synonym conflicts.

Table 5. Summary of quantitative results for RQ3. The left column shows p-values obtained via ANOVA tests for each measurement. The right column shows pairs of conditions where the effect is statistically significant or marginally significant. Significance is marked as  $p < 0.1$  (+),  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), or  $p < 0.001$  (\*\*\*).

Measurement	Book	Search Engine	Chatbot	Post-Hoc Analysis
<b>Number of Nodes</b> $F(2,272)=7.20, p < .001^{***}$	Mean = 18.67 SD = 7.71	Mean = 21.21 SD = 9.34	Mean = 23.76 SD = 10.00	Chatbot > Book ***
<b>Number of Edges</b> $F(2,272)=4.13, p < .05^*$	Mean = 25.67 SD = 15.42	Mean = 30.64 SD = 18.54	Mean = 32.83 SD = 17.45	Chatbot > Book *
<b>Immediate Closed-book Test Score</b> $F(2,272)=9.59, p < .001^{***}$	Mean = 4.83 SD = 1.67	Mean = 5.55 SD = 1.27	Mean = 5.80 SD = 1.55	Chatbot > Book *** Search Engine > Book **

### 5.3 RQ3: How does the information throughput and accuracy of knowledge acquired with LLM-based chatbots compare to that acquired with traditional search methods?

To evaluate immediate knowledge acquisition, we examined a total of 276 concept maps created after the 40-minute study phase, focusing on structural metrics such as the number of nodes and edges. Across all concept maps, the number of nodes ranged from 7 to 54, and the number of edges ranged from 2 to 122. Figure 6 presents examples of digitized concept maps created by students during the study. As shown in Table 5, the average number of nodes was highest in the Chatbot condition ( $M = 23.76, SD = 10.00$ ), followed by the Search Engine condition ( $M = 21.21, SD = 9.34$ ) and then the Book condition ( $M = 18.67, SD = 7.71$ ). Similarly, the average number of edges showed the same pattern, with the Chatbot condition including the largest number of edges ( $M = 32.83, SD = 17.45$ ), followed by the Search Engine condition ( $M = 30.64, SD = 18.54$ ) and the Book condition ( $M = 25.67, SD = 15.42$ ). A significant difference was observed between the Chatbot and Book conditions for both nodes ( $p < .001$ ) and edges ( $p < .05$ ). We found that this difference aligns with the earlier finding that students completed more search Q&A pairs in the Chatbot condition.

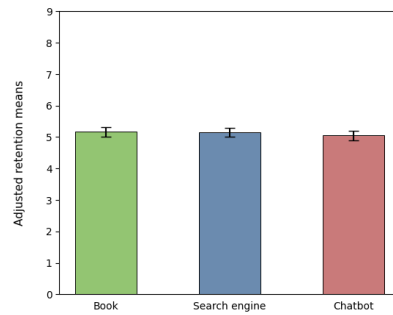


Fig. 7. Adjusted retention means across the conditions, estimated using ANCOVA with the closed-book test score as a covariate.

Table 6. Summary of quantitative results for the closed-book and two-week delayed retention tests across the three Bloom's taxonomy levels (Understand, Apply, Analyze). Values represent mean scores (0–3 scale per level), with standard deviations shown in parentheses. ANOVAs (closed-book test) and ANCOVAs (retention test) were conducted at each level across conditions. Significance is marked as  $p < 0.05$  (\*), or  $p < 0.001$  (\*\*\*).

	Closed-book test			Retention test		
	Understand	Apply	Analyze	Understand	Apply	Analyze
<b>Book</b>	1.58***, *(0.85)	1.58(1.03)	1.66(0.83)	1.62(0.88)	1.57(0.98)	1.67(0.74)
<b>Search engine</b>	1.92*(0.76)	1.79(0.92)	1.84(0.72)	1.89(0.85)	1.66(0.93)	1.69(0.76)
<b>Chatbot</b>	2.07*** (0.80)	1.84(0.92)	1.90(0.77)	1.85(0.78)	1.66(0.89)	1.76(0.78)

To further evaluate the accuracy of acquired knowledge, we examined the closed-book test scores administered right after the concept map drawing (see last row of Table 5). The average scores (scored on a 0 to 9 scale, because there were nine MCQs per condition) were highest in the Chatbot condition ( $M = 5.80, SD = 1.55$ ), followed by the Search Engine condition ( $M = 5.55, SD = 1.27$ ), and the Book condition ( $M = 4.83, SD = 1.67$ ). Scores in the Book condition were significantly lower than in both the Chatbot ( $p < .001$ ) and Search Engine conditions ( $p < .01$ ), reflecting the limited search throughput within the given time constraint.

To examine whether learning gains differed across levels of cognitive processing, we analyzed the closed-book scores by Bloom's taxonomy—Understand, Apply, and Analyze—and compared performance across conditions at each level Table 6. Although the overall average scores (scored on 0–3 scale) were highest in the Chatbot condition, followed by Search Engine and Book, modality effects were only significant at the Understand level (Chatbot>Book:  $p < .001$ ; Web>Book:  $p < .05$ ). For the Apply and Analyze levels, no significant differences were observed across conditions.

Overall, we observed that the Chatbot condition outperformed the conventional tool conditions on immediate learning outcomes, which contrasts with the perceived risk of the unreliability of LLM outputs. Rather, these empirical results provide strong evidence for students' preference for LLMs as their primary SAL tools in self-directed learning environment. However, this benefit was observed mainly at the level of recognizing facts and grasping the meanings of key concepts.



#### 5.4 RQ4: How does the long-term retention of knowledge acquired with LLM-based chatbots compare to that acquired with traditional search methods?

To examine differences in retention across conditions, we conducted an ANCOVA (Figure 7). The closed-book test score was included as a covariate and significantly predicted retention performance ( $F(1, 272) = 96.78, p < .001$ ), indicating that higher immediate learning gains were associated with higher retention scores. After adjusting for the covariate, the adjusted retention scores showed a descriptive pattern (Book : $M = 5.16$ , Search Engine : $M = 5.15$ , Chatbot : $M = 5.05$ ), but no significant differences were found across the three conditions. Similarly, ANCOVA conducted on retention test scores at each cognitive level—Understand, Apply, and Analyze—revealed no significant differences across conditions (Table 6).

Taken together, while modality effects were observed in immediate learning performance, such measurable impact was not observed in the retention test scores. The Chatbot and Search Engine conditions supported more efficient short-term learning outcomes, yet these benefits did not persist into the delayed retention test. Conversely, the Book condition produced smaller initial gains, but retention remained comparable across conditions.

## 6 DISCUSSION

### 6.1 Pedagogical Tension on Efficiency vs. Knowledge Retention in LLM-Based Chatbots for SAL

While both educators and students recognized the trade-offs of LLM use in the SAL contexts, their priorities diverged. Students valued efficient encoding, whereas educators emphasized internalization of knowledge. Our empirical results provide a complicated perspective that includes counterintuitive insight. While LLM-based chatbots facilitated significant short-term gains, these benefits were confined to lower-order thinking and did not translate into superior long-term retention compared to the book condition. We found that LLM-based chatbots enable immediate conceptual understanding, directly supporting students' preference for streamlined and targeted learning experiences. This pursuit of efficiency is particularly salient given the substantial time demands of STEM curricula, where students often dedicate numerous hours weekly to coursework [48]. Consequently, the ability of LLM-based chatbots to rapidly deliver pre-synthesized knowledge becomes an almost indispensable resource, pushing students towards greater reliance on these tools for time-sensitive tasks.

However, despite this perceived efficiency, our within-subjects experiment demonstrated a crucial discrepancy: the short-term gains associated with chatbots did not translate into higher-order reasoning nor long-term retention. While the chatbot condition optimized immediate search speed and learning outcomes, allowing students to acquire a larger number of concepts with more inter-conceptual links, the educational benefits reflected surface-level understanding and tended to decay over time. These findings necessitate a cautious approach to integrating LLM-based chatbots into SAL, urging a focus on design strategies that balance efficiency with the imperative for active, effortful, and reflective learning processes to foster robust, long-term knowledge construction.

### 6.2 Interpreting the Cognitive Efficiency and Knowledge Retention across SAL Modalities

Prior work [7, 29, 65, 66, 76, 77] consistently show that active information processing, often associated with a higher cognitive burden, contributes significantly to enhanced self-efficacy, improved learning efficiency, and, crucially, superior long-term memory formation. This phenomenon is often attributed to the principle of “desired difficulty”, where challenging yet manageable learning tasks lead to deeper processing and more robust knowledge structures [8].

The inherent interaction required by traditional books, such as manually navigating tables of contents and indices, cross-referencing concepts, and synthesizing information across chapters, inherently provides this necessary cognitive burden. Unlike the pre-synthesized output of LLM-based chatbots, the physical and structured nature of books compels learners to construct their understanding actively, leading to more profound and durable cognitive connections. This active engagement is believed to prevent the passive acceptance of information, instead fostering critical thinking and deeper processing, which are foundational for effective knowledge retention.

Furthermore, the perceived authority and reliability of books likely play a psychological role in this retention advantage. Textbooks, being curated and vetted by experts, carry a strong sense of trustworthiness. This perceived high information quality (as discussed in [Section 3.2.1](#)) likely encourages learners to fully invest in internalizing the content. In contrast, LLM-based chatbots, despite their efficiency, carry a known risk of hallucination and generally lack the same authoritative psychological weight. We speculate that this lower perceived authority or potential unreliability of chatbot-generated content may inadvertently prevent learners from fully committing information to long-term memory, as they may unconsciously (or consciously) hold back full trust in the content's veracity. Consequently, learning might remain at a more superficial level, less integrated into long-term knowledge retention, due to an underlying lack of absolute confidence in the content's reliability.

Meanwhile, the web search condition demonstrated comparable search efficiency and short-term learning gains to LLM-based chatbots. Web search requires moderate levels of learner agency, such as iterative query refinement, evaluating source credibility, and triangulating information, which introduces cognitive effort beyond passive response consumption, yet still less than the structured navigation. Recently, the boundary between web search and LLM-based chatbot interaction is becoming increasingly blurred, as web search engines integrate LLM-powered summarization features, and chatbots increasingly incorporate retrieval-augmented browsing capabilities. We note that web search was highly preferred by both educators and students in our survey studies, while still producing measurable learning outcomes comparable to those of chatbots in the within-subject experiment. We believe that this positions web search as a promising middle-ground modality. We acknowledge that familiarity may have influenced this performance advantage, as web search represents the tool users reported using most frequently for academic inquiry [[10](#), [68](#)]. Future research is needed to disentangle whether its performance stems from inherent modality affordances versus habitual familiarity.

While both web search and chatbot conditions yielded superior short-term learning gains, these advantages did not persist into delayed retention. We offer two interpretations of this pattern. First, this may reflect cognitive limits in how much pre-synthesized information can be encoded and consolidated, suggesting a natural ceiling on retention capacity [[73](#)]. Alternatively, it may indicate that the lack of desirable difficulty constrained deeper processing needed for consolidation. Future work should examine whether the observed decay reflects cognitive limits on retaining rapidly encoded information, and test whether introducing desirable difficulty into LLM-based SAL systems can improve long-term retention.

### 6.3 Towards Designing Efficient Short-term Gain and Reliable Long-term Retention using LLMs in SAL

Our findings do not suggest abandoning LLM-based tools in SAL. Given their undeniable efficiency and students' strong inclination towards them, the integration of LLMs into educational practices is arguably inevitable. The challenge, therefore, lies not in prohibition but in developing systematic guidelines and design principles that harness their power while mitigating the risks of shallow learning and fostering long-term retention. A primary design consideration must address the prior knowledge dependency issue raised among students. Current LLMs, by default, lack the sophisticated scaffolding mechanisms common in effective pedagogy. Unlike human instructors who adapt their guidance based

on a learner’s current understanding, general chatbots do not inherently detect a user’s knowledge level to provide tailored, scaffolded instruction. Future SAL-oriented LLM interfaces should incorporate features that proactively assess a learner’s existing knowledge, perhaps through initial querying or diagnostic interactions, and then dynamically adapt their responses. This intelligent scaffolding, possibly through advanced prompt engineering, could guide learners through a structured inquiry process, gradually increasing complexity and encouraging deeper engagement rather than simply delivering a final answer through designing “desired difficulty”.

Furthermore, our results strongly advocate for blended learning strategies [32] that intentionally combine the strengths of LLMs with those of traditional tools. While LLMs excel at rapid information throughput and initial encoding, these benefits did not carry over to long-term retention and higher-order thinking. Educational designs should, therefore, seek to integrate LLM interactions with more reflective activities. For instance, LLMs could serve as initial information gatherers or summarizers, and then these outputs should be brought into a classroom setting for critical discussion, peer review, or deeper analysis using traditional resources. This “flipped classroom[1]”, where AI supports initial content acquisition, could free up valuable in-class time for collaborative problem-solving, debates, and instructor-led activities designed to foster the deeper conceptual understanding and critical thinking necessary for long-term retention.

#### 6.4 Limitations

Our study has several limitations. First, it was not conducted longitudinally, which may have restricted our ability to capture more natural usage patterns in a real-world educational setting. Second, the controlled experiment was limited to a university-level STEM course with structured, fact-based content, which may have reduced the likelihood of encountering hallucination or misinformation compared to subjects involving complex interpretations (e.g., humanities) or rapidly evolving knowledge domains (e.g., semiconductor processing). Third, the participant pool was restricted to university students, which may limit the generalizability of the findings to learners with different goals or expertise levels. Lastly, while we focused on GPT-based LLM chatbots to examine autonomous learning and information retrieval, future work should assess the performance of LLMs across broader subject areas, include more diverse learner populations, and investigate emerging hybrid search tools that combine LLMs with traditional search engines to evaluate whether they address current limitations.

### 7 CONCLUSION

This research systematically investigated the effectiveness of LLM-based chatbots as a novel Search-as-Learning (SAL) tool, comparing them against traditional methods: textbooks and web search engines. Our large-scale study, involving both educator and student surveys alongside a within-subjects experiment, uncovered a fundamental pedagogical tension: students prioritize the immediate efficiency of LLM-based tools, while educators emphasize the importance of cognitive effort for deeper learning and knowledge retention. While survey results indicated strong perceived benefits and efficiency from LLM-powered chatbots, particularly among students, our empirical experiment revealed a crucial discrepancy. The chatbot condition indeed facilitated faster learning and allowed students to acquire a larger number of concepts with more inter-conceptual links in the short term. However, this immediate efficiency did not translate into higher-order thinking or long-term retention. These findings underscore a complex trade-off between processing efficiency and knowledge retention. The observed ‘easy come, easy go’ phenomenon suggests that while AI accelerates information access, it hits a natural ceiling on retention capacity, yielding long-term outcomes comparable to traditional methods. Rather than simply cautioning against AI adoption, future designs should aim to balance the convenience

of AI synthesis with the benefits of established tools, strategically optimizing SAL practices for efficient, robust, and effective learning experience.

## References

- [1] Gökçe Akçayır and Murat Akçayır. 2018. The flipped classroom: A review of its advantages and challenges. *Computers & Education* 126 (2018), 334–345.
- [2] Mayer Alvo and LH Philip. 2014. *Statistical methods for ranking data*. Vol. 1341. Springer.
- [3] Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*. Vol. 2. Elsevier, 89–195.
- [4] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 52–61.
- [5] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakci, and Rei Mariman. 2024. Generative ai can harm learning. *Available at SSRN* 4895486 (2024).
- [6] Mary Besterfield-Sacre, Jessica Gerchak, Mary Rose Lyons, Larry J Shuman, and Harvey Wolfe. 2004. Scoring concept maps: An integrated rubric for assessing engineering education. *Journal of Engineering Education* 93, 2 (2004), 105–115.
- [7] 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.
- [8] Elizabeth L Bjork, Robert A Bjork, et al. 2011. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society* 2, 59–68 (2011), 56–64.
- [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] Tara Brabazon. 2016. *The University of Google: Education in the (post) information age*. Routledge.
- [11] AJ Cañas, JD Novak, and FM González. 2004. Varieties of concept mapping. *Proceedings of the First International Conference on Concept Mapping* (2004).
- [12] Bradley W Carroll and Dale A Ostlie. 2017. *An introduction to modern astrophysics*. Cambridge University Press.
- [13] Eason Chen, Ray Huang, Han-Shin Chen, Yuen-Hsien Tseng, and Liang-Yi Li. 2023. GPTutor: a ChatGPT-powered programming tool for code explanation. In *International conference on artificial intelligence in education*. Springer, 321–327.
- [14] Xinyue Chen, Kunlin Ruan, Kexin Phyllis Ju, Nathan Yap, and Xu Wang. 2025. More ai assistance reduces cognitive engagement: Examining the ai assistance dilemma in ai-supported note-taking. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–29.
- [15] WJ Conover and RL Iman. 1979. *Multiple-comparisons procedures*. Technical Report LA-7677-MS. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States). doi:10.2172/6057803
- [16] Jennifer G Cromley, Joseph F Mirabelli, and Andrea J Kunze. 2024. Three applications of semantic network analysis to individual student think-aloud data. *Contemporary Educational Psychology* 79 (2024), 102318.
- [17] 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.
- [18] Paramveer S Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel Peter Robert. 2024. Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [19] 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.
- [20] Fiona Draxler, Albrecht Schmidt, and Lewis L Chuang. 2023. Relevance, effort, and perceived quality: Language learners’ experiences with AI-generated contextually personalized learning material. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2249–2262.
- [21] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koochang, Vishnupriya Raghavan, Manju Ahuja, et al. 2023. Opinion Paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International journal of information management* 71 (2023), 102642.
- [22] Dave L Edyburn. 1991. Fact Retrieval by students with and without learning handicaps using print and electronic encyclopedias. *Journal of Special Education Technology* 11, 2 (1991), 75–90.
- [23] 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.
- [24] Martin J Eppler and Jeanne Mengis. 2004. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society* 20, 5 (2004), 325–344.
- [25] William Estes. 2022. *Handbook of learning and cognitive processes*. Psychology Press.

- [26] Edward A Feigenbaum. 1959. *An information processing theory of verbal learning*. Rand Corporation.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAICoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–38.
- [31] Google. 2025. Google Search. <https://www.google.com/>. Accessed: Sep 1, 2025.
- [32] Charles R Graham et al. 2006. Blended learning systems: Definition, current trends, and future directions. *The handbook of blended learning: Global perspectives, local designs* 1 (2006), 3–21.
- [33] Anne-Wil Harzing, Joyce Baldueza, Wilhelm Barner-Rasmussen, Cordula Barzantny, Anne Canabal, Anabella Davila, Alvaro Espejo, Rita Ferreira, Axele Giroud, Kathrin Koester, et al. 2009. Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International business review* 18, 4 (2009), 417–432.
- [34] Bernard J Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management* 45, 6 (2009), 643–663.
- [35] Wilsaan M Joiner and Maurice A Smith. 2008. Long-term retention explained by a model of short-term learning in the adaptive control of reaching. *Journal of neurophysiology* 100, 5 (2008), 2948–2955.
- [36] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [37] Kenneth A Kiewra. 1989. A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review* 1, 2 (1989), 147–172.
- [38] Hoil Kim. 2023. *Sampling Methodology*. Kyungmoon Publishing.
- [39] Jin Young Kim, Henry Feild, and Marc Cartright. 2012. Understanding book search behavior on the web. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 744–753.
- [40] Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95, 2 (1988), 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>
- [41] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitsky, Iris Braunstein, and Pattie Maes. 2025. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872* (2025).
- [42] DR Krathwohl. 2002. A Revision Bloom’s Taxonomy: An Overview. *Theory into Practice* (2002).
- [43] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.
- [44] Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [45] Peter H Lindsay and Donald A Norman. 2013. *Human information processing: An introduction to psychology*. Academic press.
- [46] Gary Marchionini. 1995. *Information seeking in electronic environments*. Cambridge university press.
- [47] Microsoft. 2025. Bing Search. <https://www.bing.com/>. Accessed: Sep 1, 2025.
- [48] Jacob M. Miller. 2025. I Ran The Numbers. There is a 300% Workload Gap Between Some Majors. *The Harvard Crimson (Opinion)* (Feb 27 2025). <https://www.thecrimson.com/article/2025/2/27/miller-harvard-course-workload-divisions/>
- [49] Fengran Mo, Chuan Meng, Mohammad Aliannejadi, and Jian-Yun Nie. 2025. Conversational search: From fundamentals to frontiers in the LLM era. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 4094–4097.
- [50] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, et al. 2024. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters’ utilization and perceptions. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100027.
- [51] 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, Medellin, Colombia: Universidad EAFIT*.
- [52] 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.
- [53] Joseph D Novak. 1990. Concept mapping: A useful tool for science education. *Journal of research in science teaching* 27, 10 (1990), 937–949.
- [54] 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).

- [55] OpenAI. 2025. ChatGPT-4o. <https://openai.com/index/hello-gpt-4o> Accessed: Sep 1, 2025.
- [56] Fred GWC Paas and Jeroen JG Van Merriënboer. 1993. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human factors* 35, 4 (1993), 737–743.
- [57] Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 5–15.
- [58] Zachary A Pardos and Shreya Bhandari. 2024. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one* 19, 5 (2024), e0304013.
- [59] Sungjin Park. 2023. *Introduction to SQL and NoSQL Databases*. Saengneung Publishing.
- [60] José Quiroga Pérez, Thanasis Daradoumis, and Joan Manuel Marqués Puig. 2020. Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education* 28, 6 (2020), 1549–1565.
- [61] Michael A Peters, Liz Jackson, Marianna Papastephanou, Petar Jandrić, George Lazaroiu, Colin W Evers, Bill Cope, Mary Kalantzis, Daniel Araya, Marek Tesar, et al. 2024. AI and the future of humanity: ChatGPT-4, philosophy and education–Critical responses. *Educational Philosophy and Theory* 56, 9 (2024), 828–862.
- [62] Tung Phung, José Cambronero, Sumit Gulwani, Tobias Kohn, Rupak Majumdar, Adish Singla, and Gustavo Soares. 2023. Generating high-precision feedback for programming syntax errors using large language models. *arXiv preprint arXiv:2302.04662* (2023).
- [63] David R. Thorne. 2006. Throughput: a simple performance index with desirable characteristics. *Behavior research methods* 38, 4 (2006), 569–573.
- [64] 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*.
- [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] 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.
- [67] Amanda J Rockinson-Szapkiw, Jennifer Courduff, Kimberly Carter, and David Bennett. 2013. Electronic versus traditional print textbooks: A comparison study on the influence of university students’ learning. *Computers & Education* 63 (2013), 259–266.
- [68] Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R Jamali, Tom Dobrowolski, and Carol Tenopir. 2008. The Google generation: the information behaviour of the researcher of the future. In *Aslib proceedings*, Vol. 60. Emerald Group Publishing Limited, 290–310.
- [69] Kara Sage, Heather Augustine, Hannah Shand, Kaelah Bakner, and Sidney Rayne. 2019. Reading from print, computer, and tablet: Equivalent learning in the digital age. *Education and Information Technologies* 24, 4 (2019), 2477–2502.
- [70] Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. 2024. AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [71] Herbert A Simon. 1978. *Information-processing theory of human problem solving*. Erlbaum Hillsdale, NJ.
- [72] Amanda Spink. 1997. Study of interactive feedback during mediated information retrieval. *Journal of the american society for information science* 48, 5 (1997), 382–394.
- [73] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76.
- [74] John Sweller, Jeroen JG Van Merriënboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10, 3 (1998), 251–296.
- [75] Penny Thompson. 2013. The digital natives as learners: Technology use patterns and approaches to learning. *Computers & Education* 65 (2013), 12–33.
- [76] 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.
- [77] 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.
- [78] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.
- [79] Amber Walraven, Saskia Brand-Gruwel, and Henny PA Boshuizen. 2008. Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior* 24, 3 (2008), 623–648.
- [80] Mary Katherine Watson, Joshua Pelkey, Caroline R Noyes, and Michael O Rodgers. 2016. Assessing conceptual knowledge using three concept map scoring methods. *Journal of engineering education* 105, 1 (2016), 118–146.
- [81] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 214–229.
- [82] Stephen A Weyer. 1982. The design of a dynamic book for information search. *International Journal of Man-Machine Studies* 17, 1 (1982), 87–107.
- [83] Georgios N Yannakakis and Héctor P Martínez. 2015. Ratings are overrated! *Frontiers in ICT* 2 (2015), 13.



## A 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 STEM module, consisting of 3 questions from each of the selected taxonomy categories. These questions were designed to cover each module's learning objectives (LO). According to previous studies [19, 23], generating MCQs through large language models (LLMs) is preferable due to its time and cost efficiency, quality comparable to human-generated items, and alignment with Bloom's taxonomy; therefore, the MCQs were generated using ChatGPT-4o [55].

### A.1 Prompts

As shown in Figure 8, a text-based prompt specified the question requirements. To ensure consistent formatting, we also included images showing the exact format guidelines.

#### 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 **database**s 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

<p><b>Taxonomy:</b> Apply / <b>Qtype:</b> Fill in the blank</p> <p>In Python, to implement binary search recursively, what condition should be checked first? Fill in the blank below.</p> <pre>def binary_search_recursive(arr, x, start, end):     if ___:         mid = (start + end) // 2         if arr[mid] == x:             return mid         elif arr[mid] &gt; x:             return binary_search_recursive(arr, x, start, mid - 1)         else:             return binary_search_recursive(arr, x, mid + 1, end)     return -1</pre> <p>A) start &lt; end B) start &lt;= end C) start == end D) start &gt; end</p> <p><b>Correct Answer:</b> B / <b>Distractors:</b> Two distracting incorrect answer <b>Explanation:</b> Tell me why B is the correct answer.</p>
--

Fig. 8. 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.

### A.2 Quality Evaluation

The evaluation was conducted using two levels of metrics: question-level [19] and quiz-level [19, 23]. We initially generated three times more items than required to allow quality screening. In total, 81 MCQs were created (27 per



module), and the closed-book test containing 9 MCQs per module was finalized through a three-step refinement process. *Step 1)* Four of the authors cross-validated 81 initial questions using metrics (Table 7), and two of them further reviewed the questions for alignment with learning objectives (LO) and taxonomy, removing items that did not meet the criteria. *Step 2)* After refinement and evaluation, the highest-rated items based on the question-level metrics were selected, resulting in a closed-book test consisting of 9 MCQs per module. *Step 3)* For the quiz-level evaluation, we recruited three domain experts, all of whom were university professors. The experts evaluated the closed-book tests based on three criteria—structure, redundancy, and usefulness (Table 3). Experts received 100,000 KRW (approx.73 USD) as compensation for participation.

	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 7. 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.

### A.3 Examples of MCQs

#### Example 1) Understand

You are classifying different types of database management systems. Look at the features below and determine which type of system each feature describes.

Feature	System Type
Uses tables to store data	???
Ensures ACID properties	???
Suitable for hierarchical data storage	???
Supports complex joins	???

- A) RDBMS, RDBMS, Non-RDBMS, RDBMS
- B) Non-RDBMS, RDBMS, RDBMS, Non-RDBMS
- C) RDBMS, Non-RDBMS, RDBMS, RDBMS
- D) RDBMS, Non-RDBMS, RDBMS, Non-RDBMS

#### Example 2) Apply

In SQL, to insert a new record into the **employees** table, use the **INSERT INTO** statement. Fill in the blank with the appropriate SQL command to insert a record.

```
INSERT INTO employees (id, name, position)
_____, 'John Doe', 'Manager');
```

- A) **SELECT**
- B) **VALUES**
- C) **SET**
- D) **UPDATE**

#### Example 3) Analyze

What will be the output of the following SQL commands?

```
CREATE TABLE employees (
  id INT PRIMARY KEY,
  name VARCHAR(50),
  department VARCHAR(50),
  salary DECIMAL(10, 2)
);

INSERT INTO employees (id, name, department, salary) VALUES (1, 'John', 'HR', 60000.00);
INSERT INTO employees (id, name, department, salary) VALUES (2, 'Jane', 'Finance', 52500.00);
INSERT INTO employees (id, name, department, salary) VALUES (3, 'Bob', 'HR', 10000.00);
SELECT * FROM employees WHERE department = 'HR';
SELECT department, COUNT(*) AS employee_count, AVG(salary) AS average_salary
GROUP BY department;
DROP TABLE employees;
```

- A) Table dropped, no output.
- B) One row: (Finance, 1, 60000.00)
- C) One row: (HR, 2, 52500.00)
- D) Two rows: (HR, 1, 52500.00) and (Finance, 1, 60000.00)

Fig. 9. As part of the post-test, nine questions from the database module were used, and one example from each taxonomy is provided as follows.