Scaffolding Inquiry-Oriented Web Search using LLM-based Question Generation

Yusuke Yamamoto
Nagoya City University
Nagoya, Japan
yusuke_yamamoto@acm.org

***Abstract*—Herein, we propose a method that promotes inquiry-oriented information seeking using a search engine to critically evaluate and synthesize diverse information when deciding on topics for which no definitive answer exists. We introduce a context-aware scaffolding framework in which a virtual student and teacher – both instantiated as LLM agents – jointly generate and rank driving questions aligned to a web searcher's in-session browsing context. Unlike prior search-as-learning interventions that rely on static adjunct or pre-specified "expected" questions, our method displays questions to a web searcher in real time to the pages the searcher actually reads and selects the most pedagogically valuable question prompt, thereby deepening their opinions and providing them with new perspectives and insights. We evaluated the approach through an online user study (N=120) and a controlled laboratory study (N=24). We found that displaying scaffolding questions significantly increased the participants' dwell time on search engine result pages without inflating the overall task duration. In addition, the prompts were considered relevant, important, and helpful for organizing ideas. The intervention nudged the participants toward a question-driven search strategy that enhanced their comprehension of the inquiry topic.**

***Index Terms*—Information retrieval, Search as learning, Inquiry-oriented information seeking, Scaffolding, LLM**

I. Introduction

The emergence of large language models (LLMs) has changed the nature of information seeking and decision-making. Conversational generative artificial intelligence (GenAI) systems built on LLMs can instantly generate fluent answers to natural language queries. Thus, conversational GenAI has attracted attention as an information access tool that can serve as an alternative to web search engines.

However, there are various concerns about using conversational GenAI for some topics, such as life events or social issues, where accuracy is crucial or for which there is no single correct answer. GenAI often produces misinformation called "hallucinations" when LLM does not have the information requested [11]. Conversational GenAI produces answers that reinforce users' prior beliefs, and users tend to accept AI answers without scrutiny [17]. Therefore, for topics on which the LLM lacks knowledge or for which multiple diverse answers may exist, it is necessary for users to proactively gather, examine, and organize information and derive answers without excessively relying on conversational GenAIs.

Inquiry-oriented information seeking can be defined as the use of a search engine to evaluate and synthesize diverse information while making decisions about a topic for which there is no definitive answer. Typically, web search engines are used in isolation, and there is usually no conversational partner during a search. Therefore, search engine users cannot easily formulate questions to deepen their knowledge of unfamiliar topics. Furthermore, users avoid seeking information that does not support their pre-existing beliefs [20], [18].

In inquiry-oriented information seeking, it is necessary to actively pose questions, gain a deep understanding of the subject, and form one's perspective while examining and organizing various information. Herein, we propose a system that automatically generates and presents *scaffolding* question prompts to activate users' inquiry-oriented information seeking in response to their web search behavior. Scaffolding has been proposed to promote learners' active and deep learning activities in learning sciences [21]. Scaffolding is a teaching method in which a more capable other provides hints or clues tailored to the learner's context so that the learner can understand the subject and solve problems independently. In the proposed system, we regard a web searcher as a learner and emulate a teacher to provide scaffolding for an inquiry-oriented search using an LLM. In particular, each time a searcher views a webpage, the proposed system generates a "question prompt" to deepen the user's understanding and insight about the topic inquired based on their browsing behavior and presents it on the search results interface.

Figure 1 shows the scaffolding question presented by the proposed system for a web searcher inquiring on the topic "How to ensure learning opportunities for low-income children in urban areas." Before returning to the search engine results page, the searcher searched "educational support schemes for low-income families" on a webpage. The proposed system displayed a scaffolding question (in the lower-right corner of the screen): "How can the entire local community support the learning of children from impoverished households"? The page viewed by the searcher summarizes municipal grant programs aimed at alleviating financial burdens. If the searcher has gleaned ideas about government-led educational support from this page, the system's question – shifting the focus to the "entire local community" – may prompt reflection on the potential roles of corporations, nonprofit organizations, ordinary citizens, and municipal authorities. By generating and presenting such scaffolding questions in response to the searcher's behavior, the system can clarify the inquiry's

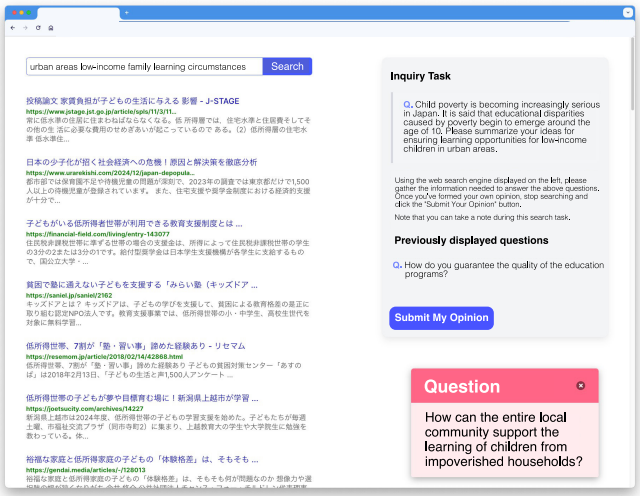

Fig. 1: Example of a scaffolding question prompt stimulating inquiry-oriented information seeking. The prompt is displayed at the bottom-right of the screen (a part of the text is translated into English for paper readers).

trajectory, broaden the range of perspectives considered, and ultimately invigorate inquiry-oriented information seeking.

To evaluate the proposed approach, we compared it to two existing approaches: (a) posing a set of predetermined questions derived solely from the assigned task [3], and (b) prompting web searchers to recall the content of the page they have just read [27]. We considered the following key questions:

- Impact on user behavior and strategies: How do context-adaptive scaffolding question prompts affect the information-seeking behavior, attitudes, and strategies of users engaged in an inquiry task?
- Quality of question prompts: How do users evaluate the relevance and usefulness of the presented scaffolding prompts?

II. Related Works

A. Enhancing information seeking behavior

In information retrieval, various interaction techniques have been studied to enhance the quality of users' web search behavior. Umemoto et al. proposed a search interface that visualizes users' coverage of subtopics necessary for understanding the search topic, thereby promoting comprehensive information exploration [19]. Yamamoto et al. proposed a query priming method that incorporates critical-thinking-stimulating words in query suggestions to encourage critical web search behavior [23]. Harvey et al. demonstrated that presenting users with examples of effective queries and the quality score of their queries on the search engine interface improves their query formulation skills [7].

Studies have also begun to explore how to enhance information-seeking behavior when interacting with conversational generative AIs. Danry et al. reported that when generating responses using conversational AI, ending with a question that prompts users to evaluate the validity of a given reason activates users' critical thinking more effectively than simply presenting the reason [5]. Lee et al. proposed conversational design guidelines for conversational AI to increase children's engagement and elicit further questions when responding to children's inquiries [9]. Yamamoto proposed a method to elicit proactive questioning behavior from users by deliberately ending conversational AI responses in a way that encourages curiosity about what comes next [22].

B. Question generation

Several studies have considered the presentation of questions to users via web search engines to enhance their search experience.

Zhao et al. proposed an algorithm that transforms users' keyword queries into question sentences using templates and presents them as candidate queries to improve the QA search ranking [26]. Rosset et al. developed a language model that generates natural language questions to clarify the user's intent from keyword-based queries entered into a search engine [16]. Zhu et al. proposed a method for generating questions related to the content of a webpage immediately after a user has viewed it to enhance users' understanding of the search topic [27].

Some studies have focused on question generation to assess or promote learning. Cheng et al. proposed a method that breaks down conceptual elements to automatically generate questions of varying difficulty levels and evaluate students' understanding of any given concept [3]. Zhang et al. proposed conversational AI that generates interesting stories about mathematical concepts and automatically engages in Q&A to effectively teach children [25]. Our scaffolding-question-prompting method proposed herein aims to encourage users' reflection for inquiry-oriented search and can be considered a form of follow-up question generation to support search activities.

C. Search as learning

Information seeking to understand a topic is conceptualized as *search as learning*. Search as learning can be classified into three types according to the user's learning (cognitive level) and information-seeking modes: look-up, exploratory, and comprehensive search [15]. A look-up search involves seeking specific factual information. The user learns factual knowledge by understanding and memorizing the retrieved information. Exploratory search involves evaluating, analyzing, and integrating multiple pieces of information to understand a given topic. A comprehensive search is similar to an exploratory search in its integration of information; however, it aims at acquiring new insights or perspectives from the learned content. The inquiry-oriented information seeking targeted herein can be regarded as a comprehensive search.

At the intersection of information retrieval and human-computer interaction, several studies have explored technologies that support search as learning. Yu et al. proposed a method for estimating users' knowledge gain on a topic from

web search behavior, such as query logs, page dwell time, and result-page viewing time [24]. Câmara et al. analyzed the effect of presenting structured subtopics to users learning about a theme using a web search engine [2].

Câmara focused on supporting understanding and memory retention in look-up and exploratory searches. In contrast, we focus on searcher-AI interactions that promote the integration and organization of collected information and the formation of opinions in a comprehensive search.

III. Proposed Method

This section describes the "scaffolding" question generation method designed to promote inquiry-oriented information seeking aligned with web-search behavior.

In general, users' search behavior is observable; however, their internal thoughts and opinions during searches are inaccessible. To address this, the proposed approach emulates a virtual student using an LLM who performs the same task as the user.

Each time a user views a webpage, the system estimates the opinion that a good student might form upon viewing that same page. Subsequently, a virtual teacher, also emulated with an LLM, generates candidate questions to guide the user toward the inferred opinion. The virtual student then selects the best question. The selected question is presented to the user to facilitate inquiry-oriented information seeking. The proposed approach is described in detail below.

A. Generating a virtual student opinion

In the first step, it is assumed that a virtual student, emulated by an LLM, is working on the same inquiry task as a web search user. This student generates an opinion statement that serves as a basis for question generation.

In particular, each time a user views a webpage, it is assumed that the virtual student agent simultaneously views the same page. The system then generates a text that includes the student's opinion on the task theme and the reasons for that opinion based on the content of the viewed page. If the student agent already holds an opinion, it updates the opinion by considering the newly viewed page and any questions previously presented to the user. After generating an opinion, the system stores it internally along with the page content and a list of previously presented questions. At the start of the inquiry task, the student agent has no prior opinion, and the page view history and question list are empty.

Figure 2 shows the prompt for generating a virtual student's opinion using LLM. The prompt explicitly restricts the use of knowledge other than that provided in the fields.

B. Generating a list of questions leading to model answers

In the next step, a list of candidate questions that activate the inquiry is generated based on the opinion produced by the virtual student agent in the previous step. To achieve this, a virtual teacher is emulated using an LLM that performs the reverse generation of questions whose ideal answer would be the student's opinion.

You are a good student. Your teacher tells you to summarize your opinion for "Theme." Assume that you currently hold "Current opinion" and then learn "Relevant information." Revise your current opinion and provide a new opinion on the "Theme". Please make sure to follow the "Constraints."

=== Theme ===
{Task theme}

=== Current opinion ===
{Prior opinion}

=== Relevant information ===
{Content of most recently viewed page}

=== Supplemental questions ===
{List of questions previously displayed to the user}

=== Constraints ===
(1) Only use the information in "Current opinion" and "Relevant information." (2) Include reasoning and supporting evidence in your opinion. (3) If "Supplemental questions" are present, take them into account when forming your opinion.

Fig. 2: Prompt for generating a virtual student's opinion.

You are a good teacher. Your task is to generate N good questions that help students learn about a given subject. Example pairs of good questions and their corresponding learning contents are listed below. Based on these examples, generate a good question list for the specified "Learning content."

=== Learning content ===
{Opinion generated by virtual student in previous step}

=== Examples ===
Learning content: {Answer in few-shot examples}
Question: {Question in few-shot examples}

Fig. 3: Prompt for generating a list of candidate questions.

Figure 3 shows the prompt used to generate a list of questions. We apply few-shot prompting by including examples of high-quality question-answer pairs to guide the generation process.

C. Selecting a driving question

Finally, the most effective question from the generated questions is selected to activate inquiry-oriented information seeking. In learning sciences, questions that provide a context

You are currently learning about the "Theme." You hold the "Opinion." Your teacher has presented you with "Question list." Rank these questions, following the "Constraints."

=== Theme ===
{Task theme}

=== Opinion ===
{Opinion generated by virtual student in step 1}

=== Constraints ===
(1) Follow the "Criteria" when assessing questions. (2) The more criteria a question satisfies, the better its quality.

=== Criteria ===
(1) Promotes higher-order thinking (2) Appropriately complex and can be decomposed into subquestions (3) Connects concepts across disciplines (4) Can be answered through investigation (5) Related to real-world situations (6) Relevant to the learner's everyday life (7) Personally interesting and intellectually stimulating (8) Related to the "Theme" (9) Not similar to the "List of previously asked questions"

=== List of previously asked questions ===
{Questions already displayed to a user by the system}

=== Question list ===
{Questions generated by a virtual teacher in step 2}

Fig. 4: Prompt for ranking the list of driving questions.

for pursuing inquiry toward a goal and impart coherence and continuity to the entire inquiry process are referred to as "driving questions" [8]. Well-crafted driving questions can motivate learners and help them recognize the importance of a problem that needs to be solved [14].

The proposed method employs LLM to emulate a virtual student who evaluates the quality of each candidate question as a driving question. Figure 4 shows the prompt used to rank the candidate questions. The evaluation criteria proposed by Krajcik et al.[1] are adopted, and additional criteria are considered to ensure that questions irrelevant to the theme or similar to previously shown questions are not highly ranked. The system ranks questions based on these criteria and presents the highest-ranked question to the user.

D. Research questions

In inquiry-oriented information seeking, success can be evaluated from two perspectives: the quality of the outcomes (e.g., opinions formed from the task) and the information-seeking process.

[1]https://websites.umich.edu/~krajcik/DQ.html

A previous study on search as learning evaluated the effectiveness of learning via look-up or exploratory search by analyzing task outcomes submitted by participants to assess their conceptual understanding [12], [2], [27]. However, inquiry-oriented information seeking involves exploring open-ended topics without definitive answers; thus, it requires a critical evaluation and synthesis of diverse information. Therefore, defining a "correct" answer and evaluating task success solely by analyzing the outcome are inappropriate. In addition, the quality of outcomes in an inquiry-oriented search depends not only on the users' strategies and attitudes but also on high-level cognitive skills, such as evaluating, organizing, integrating, and generating information. Improving such advanced cognitive skills in a short term is difficult; thus, they are difficult to assess through short-term experiments.

On the other hand, users' attitudes and search strategies during the information-seeking process can be influenced in the short term by system interventions. Changes in these processes can be evaluated by analyzing the search behavior and user thought processes during the task.

Based on this consideration, this study focuses on the impact of scaffolding questions on information-seeking processes. We consider the following research questions:

RQ1 How do scaffolding questions affect the search behavior of users engaged in inquiry-oriented tasks?
RQ2 How do users who are engaged in inquiry-oriented information seeking evaluate the quality of scaffolding questions presented to them?
RQ3 How do scaffolding questions influence users' information-seeking attitudes and strategies during inquiry tasks?

IV. User Study 1: Online Experiment

We conducted an online study using a crowdsourcing platform in Japan to examine **RQ1** and **RQ2**. The study involved 120 participants and was conducted from November 14 to 15, 2024. We employed a between-subjects design and compared four prompting conditions (UI conditions).

A. Procedure

After providing consent, the participants were redirected to our website for the online study. The participants read a task instruction and then rated their prior knowledge on four candidate task themes. Each participant was assigned the theme with the lowest self-reported knowledge and randomly allocated to a UI condition.

Next, the participants read the following inquiry-oriented search task scenario related to their assigned theme:

> *Child poverty is becoming increasingly serious in Japan. It is said that educational disparities caused by poverty begin to emerge around the age of 10. Please summarize your ideas for ensuring learning opportunities for low-income children in urban areas.*

After reading the scenario, the participants used our web search engine to collect information for the task. An initial

TABLE I: Themes of the inquiry-oriented search tasks and participants' prior knowledge (measured on seven-point Likert scales). The values in parentheses under the theme column indicate the number of participants. The knowledge is presented as means and standard deviations.

Theme	Knowledge
How to resolve conflicts with foreign apartment residents (42)	1.57 (0.80)
How to prevent infectious diseases in developing countries (30)	1.60 (0.77)
How to protect impoverished households from disasters (26)	1.46 (0.76)
How to ensure learning access for low-income children (16)	1.31 (0.49)

TABLE II: Example prompts shown under each UI condition when a user explores the theme "How to ensure learning access for low-income children" and views a page about municipal support programs.

UI condition	Example of question prompt
SCAFFOLDING	How can local communities collectively support the education of children from low-income families?
EXPECTED	What kinds of educational opportunities are often inaccessible to children in urban low-income households?
ADJUNCT	What is the difference between the School Assistance Program and High School Tuition Support Fund?

query, prespecified by the authors, was entered into the search box to help participants broadly explore the theme (e.g., "urban areas low-income family learning circumstances"). There was no time limit for this information-seeking phase. When the participants felt they had collected sufficient information, they clicked the "Submit My Opinion" button at the top-right corner of the search interface (Figure 1) and proceeded to an opinion-writing page. On this page, the participants were asked to freely describe their opinions on the assigned theme. Under the UI conditions in which question prompts were shown, a list of previously presented prompts was also displayed as reference information.

After submitting their opinions, the participants completed a post-questionnaire, including the evaluation of the question prompts displayed during the web search and demographic questions.

B. Task themes

We selected the task themes according to the following criteria: (1) Unfamiliar to the participants, (2) Do not require highly specialized knowledge to understand, (3) Require synthesizing information from multiple perspectives, (4) Do not have a definitive answer, allowing for creative solutions

We collected themes that met these criteria from previous university entrance essay questions in the social science. We slightly modified them and obtained the four final themes (Table I).

C. UI conditions

We compared four UI conditions: a baseline condition simulating a standard web search engine (PLAIN) and three extended conditions incorporating question prompts – SCAFFOLDING, EXPECTED, and ADJUNCT. For the three extended UIs, the system presented generated questions in the bottom-right corner of the SERP screen whenever participants returned from previously visited webpages (Figure 1). Table II lists example question prompts generated under each condition. These examples assume that a user is exploring the theme "How to ensure learning opportunities for low-income children in urban areas" and has just viewed a page about municipal support programs.

*Standard web search UI (*PLAIN *condition):* The PLAIN condition simulates a conventional web search interface, like Google. Typing a query into the search bar returns a list of web search results. We used the Microsoft Bing Web Search API[2] to retrieve the search results (limited to 50 items per query). Each result included a title, URL, and summary. To keep the task theme visible at all times, the right sidebar of the search interface displayed the assigned inquiry theme and included a button linked to the opinion writing page.

*Scaffolding question prompt UI (*SCAFFOLDING *condition):* The SCAFFOLDING UI extends the PLAIN UI to incorporate scaffolding question prompts generated using the method in Section III.

When a participant remains on a webpage for more than three seconds, the system interprets the page as "read" and summarizes its content using an LLM, subsequently caching the summary. A new question prompt is then generated based on the task theme, latest page summary, current opinion of the virtual student agent (initialized as "none" at the start), and previously displayed question prompts. Summarization and question generation were performed using the Google Gemini API (Gemini 1.5 Flash)[3]. Few-shot prompting was employed using 10 example QA pairs from the ELI5 dataset[4], which comprises questions requiring elaborate answers.

*Predefined question prompt UI (*EXPECTED *condition):* The EXPECTED condition also extends the PLAIN UI but displays predefined question prompts during the search task. This approach is similar to a teacher-driven instructional model in which necessary questions are prepared in advance and shown to learners.

For each task theme, the Google Gemini API was used to generate 20 question prompts before the experiment. The input prompt to the Gemini API was "*You are a skilled teacher. Generate 20 questions that students should consider in order to answer the following theme: {Theme}*". Whenever a participant viewed a webpage for more than 3 sec and returned to the search interface, an unused question prompt was randomly selected from the 20 and shown in the bottom-right corner of the screen.

*Adjunct question prompt UI (*ADJUNCT *condition):* The ADJUNCT condition also builds upon the PLAIN UI and follows the method proposed by Zhu et al. [27], which generates questions from webpages such that the webpages contain ideal answers for the questions using a BART model fine-tuned on the ELI5 dataset [10].

[2]Bing Web Search API: https://www.microsoft.com/en-us/bing/apis/
[3]Google Gemini API: https://ai.google.dev/gemini-api/docs
[4]ELI5 Dataset: https://facebookresearch.github.io/ELI5/explore.html

We implemented a simplified version of the method using the Google Gemini API. When a participant viewed a webpage for more than 3 sec, the system generated and suggested a question from the summarized content of the page using few-shot prompting (with the same example 10 QA pairs as the SCAFFOLDING).

D. Participants

We recruited 120 participants through CrowdWorks[5], a major Japanese crowdsourcing platform. We excluded six participants due to inattentive responses or system errors; thus, we had 114 valid participants. Among them, 67 were male, 46 were female, and 1 did not disclose their gender. The median age group was 40s (20s: 12.3%, 30s: 32.5%, 40s: 28.1%, 50s: 22.8%, 60s: 3.5%, NA: 0.8%). For educational background, 62 participants had a university or graduate degree, and 52 did not. Participants were randomly assigned to one of the four UI conditions: PLAIN (N=30), ADJUNCT (N=34), EXPECTED (N=26), and SCAFFOLDING (N=24). Each participant received 600 JPY (approximately 4 USD) upon completing the task. The average task duration was 29.5 min.

E. Metrics

Search behaviors: User behavior during web searches was considered to implicitly reflect the cognitive processes involved in inquiry-oriented information seeking. Based on prior studies on search as learning [4], [2], [27], we collected search behavior logs focusing on the following metrics: (1) search task duration, (2) total SERP duration, (3) average SERP duration per query, (4) number of queries issued, and (5) number of clickthroughs.

Semantic relevance between prompting questions and opinion statements: If prompting questions influence inquiry-oriented information-seeking tasks, their impact should be evident in not only in search behavior but also the outcomes of the tasks. Therefore, for the three UI conditions that presented questions (SCAFFOLDING, EXPECTED, and ADJUNCT), we calculated the semantic relevance between the prompting questions displayed during the web searches and the opinion statements reported by the participants.

Let $Q_u = \{q_1, ..., q_n\}$ denote a set of questions presented to participant u during the search task, and o_u represents the opinion submitted by user u. The relevance $Rel(Q_u, o_u)$ can be expressed as follows:

$$Rel(Q_u, o_u) = \frac{1}{|Q_u|} \sum_{q \in Q_u} sim_{cos}(v(q), v(o_u)) \qquad (1)$$

where $v(t)$ is the embedding vector of text t, and $sim_{cos}(v_1, v_2)$ is the cosine similarity between the vectors v_1 and v_2. A higher value of $Rel(Q_u, o_u)$ indicates that the opinion statement is more relevant to the presented question prompts. To vectorize the question prompts and opinion statements, we employed the Japanese sentence embedding model GLuCoSE v2[6].

[5]CrowdWorks: https://crowdworks.jp/

[6]GLuCoSE v2: https://huggingface.co/pkshatech/GLuCoSE-base-ja-v2

Quality of prompting questions: For the SCAFFOLDING, EXPECTED, and ADJUNCT conditions, participants rated each prompting question displayed during web searches based on five-point Likert items: relevance, importance, interest, answerability, and usefulness for deepening understanding of the theme, gaining new perspectives, and organizing information.

F. Results

We conducted statistical analyses on the collected data to evaluate the effects of each UI condition. Because the data did not satisfy the normality assumption, the Kruskal-Wallis test, a non-parametric method for variance analysis, was employed for hypothesis testing. For the post-hoc analysis, we applied the Benjamini-Hochberg procedure to control the false discovery rate (FDR) in multiple comparisons among the UI conditions [1]. The significance level of all hypothesis tests was set to 0.05. Table III presents the means and standard deviations of each metric for the online experiment.

1) Search behavior metrics: As presented in Table III, the total and average SERP duration per query significantly varied across the UI conditions.

The mean total SERP durations for the PLAIN, ADJUNCT, EXPECTED, and SCAFFOLDING conditions were 145.4, 264.4, 291.3, and 285.3 s, respectively. The Kruskal–Wallis test revealed significant differences among the conditions ($p < 0.05$). Post-hoc pairwise comparisons using the Benjamini-Hochberg adjusted Mann-Whitney U tests revealed that the EXPECTED condition had a significantly longer SERP duration than the PLAIN condition ($p < 0.05$).

The mean average SERP durations per query for the PLAIN, ADJUNCT, EXPECTED, and SCAFFOLDING conditions were 59.3, 95.3, 96.0, and 104.8 s, respectively. The Kruskal-Wallis test confirmed significant differences among the conditions ($p < 0.05$). The ADJUNCT and SCAFFOLDING conditions exhibited significantly longer average SERP durations per query than PLAIN ($p < 0.01; p < 0.01$).

Notably, under the SCAFFOLDING, EXPECTED, and ADJUNCT conditions, prompting questions were displayed on the SERPs. These results demonstrate that participants under these conditions may have experienced cognitive engagement induced by the question prompts, resulting in longer dwell times on the SERP pages.

2) Relevance between questions and opinions: To assess the influence of the prompting questions on opinion formation, we compared the semantic relevance between the question prompts displayed during the task and the participants' final opinion statements across the UI conditions. The mean semantic relevance scores for ADJUNCT, EXPECTED, and SCAFFOLDING conditions were 0.667, 0.710, and 0.723, respectively. The Kruskal-Wallis test revealed significant differences ($p < 0.001$). Post-hoc tests revealed that both EXPECTED and SCAFFOLDING conditions had significantly higher relevance scores than ADJUNCT ($p < 0.05; p < 0.01$). These results suggest that participants under the EXPECTED and SCAFFOLDING conditions were more likely to produce opinions related to the prompts than those under the ADJUNCT condition.

TABLE III: Means and standard deviations of each metric for the online user study. Asterisks (*) indicate metrics for which significant differences across UI conditions were found in the variance analysis. Superscripts $\mathcal{P}$, $\mathcal{A}$, $\mathcal{E}$, and $\mathcal{S}$ indicate statistically significant differences in pairwise comparisons with PLAIN, ADJUNCT, EXPECTED, and SCAFFOLDING conditions, respectively.

	UI Condition			
Metric	PLAIN	ADJUNCT	EXPECTED	SCAFFOLDING
Search behavior metrics				
Task duration (s)	745.8 (476.8)	833.4 (586.1)	869.0 (636.9)	799.0 (480.8)
Total SERP duration (s)*	145.4 (120.8)$^{\mathcal{E}}$	264.4 (240.1)	291.3 (256.7)$^{\mathcal{P}}$	285.3 (297.4)
Avg. SERP duration per query (s)*	59.3 (95.4)$^{\mathcal{AS}}$	95.3 (66.8)$^{\mathcal{P}}$	96.0 (95.4)	104.8 (86.5)$^{\mathcal{P}}$
Number of queries	3.07 (2.42)	3.09 (2.49)	3.81 (2.42)	3.08 (2.43)
Number of clickthroughs	6.17 (4.79)	5.94 (3.96)	5.92 (3.33)	6.67 (4.07)
Task outcome				
Semantic relevance between prompts and opinion*	–	0.677 (0.060)$^{\mathcal{ES}}$	0.710 (0.038)$^{\mathcal{A}}$	0.723 (0.052)$^{\mathcal{A}}$
Evaluation of prompts				
Relevance to theme*	–	2.59 (2.05)$^{\mathcal{ES}}$	3.04 (2.16)$^{\mathcal{A}}$	3.15 (2.07)$^{\mathcal{A}}$
Importance for the theme*	–	2.46 (2.02)$^{\mathcal{ES}}$	2.96 (2.12)$^{\mathcal{A}}$	3.09 (2.02)$^{\mathcal{A}}$
Interestingness*	–	2.40 (1.99)$^{\mathcal{S}}$	2.88 (2.06)	2.98 (1.95)$^{\mathcal{A}}$
Ease of answering	–	2.59 (2.06)	2.68 (2.03)	2.80 (1.88)
Usefulness for deepening understanding*	–	2.39 (1.96)	2.84 (2.03)	2.94 (1.90)
Usefulness for gaining new perspectives*	–	2.22 (1.85)$^{\mathcal{S}}$	2.65 (1.95)	2.83 (1.90)$^{\mathcal{A}}$
Usefulness in organizing information/opinions*	–	2.31 (1.90)$^{\mathcal{ES}}$	2.81 (2.05)$^{\mathcal{A}}$	2.82 (1.85)$^{\mathcal{A}}$

3) Quality of prompting questions: During the online task, 398 prompting questions were presented across the 3 prompt-enabled UI conditions (SCAFFOLDING, EXPECTED, and ADJUNCT; average of 4.74 prompts per participant). To evaluate the quality of the prompts, we compared the participants' ratings for each prompt across the three conditions.

Hypothesis testing revealed significant differences across the conditions for the following items: "relevance to the theme" ($p < 0.01$), "importance for the theme", "interestingness" ($p < 0.05$), "usefulness for gaining new perspectives" ($p < 0.05$), and "usefulness for organizing collected information or opinions" ($p < 0.05$).

For relevance to the theme, post-hoc tests confirmed that both SCAFFOLDING and EXPECTED received significantly higher ratings than ADJUNCT ($M = 3.15, 3.04$, and 2.59 for SCAFFOLDING, EXPECTED, and ADJUNCT, respectively, where M is the mean value; $ps < 0.05$). In addition, the importance of question prompts under the two conditions was significantly higher than that under ADJUNCT ($M = 3.09, 2.96$, and 2.46, respectively; $ps < 0.05$). For interestingness and usefulness for gaining new perspectives, SCAFFOLDING got significantly higher ratings than ADJUNCT (interestingness: $M = 2.98$ and 2.40, and $p < 0.05$; usefulness for gaining new perspectives: $M = 2.83$ and 2.22, and $p < 0.05$). For usefulness in organizing collected information or opinions, the ratings for ADJUNCT, EXPECTED, and SCAFFOLDING were 2.31, 2.81, and 2.82, respectively. Post hoc tests showed that both SCAFFOLDING and EXPECTED scored significantly higher than ADJUNCT ($ps < 0.05$).

These results demonstrate that prompting questions under the EXPECTED and SCAFFOLDING conditions were more relevant to the inquiry theme, more important for opinion formation, and more helpful in organizing participants' collected information and ideas than those under the ADJUNCT condition. Furthermore, prompts under the SCAFFOLDING condition were perceived as more interesting and more likely to offer new perspectives for forming opinions than those under the ADJUNCT condition.

V. USER STUDY 2: LABORATORY EXPERIMENT

Based on the online study, we conducted a lab experiment to qualitatively evaluate participants' perceptions of the quality of the question prompts (**RQ2**) and the effects of prompts on attitudes and strategies during inquiry-oriented search (**RQ3**).

This study was conducted from January 20 to 31, 2025. We recruited 24 students majoring in data sciences at the authors' university as participants. One participant was excluded from the analysis due to a system error. The remaining 23 participants were randomly assigned to one of the four UI conditions: PLAIN (N=6), ADJUNCT (N=5), EXPECTED (N=7), and SCAFFOLDING (N=5). Each participant received a 2,000 JPY (approximately 13 USD) after the task. The average task duration was 76 minutes.

A. Procedure

After providing consent, the participants worked individually in a university lecture room. The participants were briefed, after which they accessed the experimental website and began the tasks using a pre-configured PC with the memo app. The tasks proceeded in the following order: (1) inquiry-oriented search task and (2) post-task questionnaire. For this experiment, we selected "How to ensure learning opportunities for low-income children" as the task theme, for which online study participants reported the least prior knowledge. The UI assignment and system behavior were consistent with the online user study. The participants were encouraged to take notes with the memo app while performing the tasks.

In addition to the post-task items used in the online study, we included an open-ended questionnaire asking participants to reflect on their web search behavior and how they used the displayed question prompts. We asked:

> *While gathering information for the task, what did you consciously focus on? Please freely describe your approach for selecting search terms, clicking web search results, and reading webpages.*

TABLE IV: Coding results on the usage, frequency, and negative perceptions of prompting questions. The numbers indicate the number of participants who mentioned each code.

Code	UI Condition		
	ADJUNCT (N=5)	EXPECTED (N=7)	SCAFFOLDING (N=5)
Question usage			
Question-driven search	1	1	4
Awareness of unexpected perspectives	2	5	1
Trigger for deeper exploration of specific viewpoints	2	0	2
Aid for organizing collected information or opinions	0	2	2
Cue to recall previously read information	2	0	0
Question usage frequency			
Actively considered the questions	1	3	3
Referred to questions as needed	3	2	0
Used questions when stuck	1	0	0
Rarely used the questions	0	0	1
Negative perceptions of questions			
Irrelevant to personal interests	2	1	0
Unrelated to the theme	1	1	0
Contradicted personal opinion and caused confusion	0	1	0

The participants assigned to the ADJUNCT, EXPECTED, and SCAFFOLDING conditions were also administered an open-ended questionnaire that asked them to reflect on the prompting questions that appeared during the task. We asked:

> *Recall that "question prompts" appeared while you were viewing a list of web search results. What did you think when these prompts were displayed? Please freely describe how the prompts might have influenced your information-seeking approach, opinion formation, or interpretation of the webpages you read.*

When responding to the open-ended questions, participants were provided with a list of all question prompts displayed during their tasks as a reference. They were encouraged to consult both the provided list and the notes they had written during the search task.

B. Results

We conducted an inductive coding analysis [6] of the open-ended questionnaire responses to analyze participants' cognitive processes during information seeking and their use of question prompts. The coding was performed using NVivo 15. To ensure reliability, a second round of coding was conducted a few days after the first. The intracoder reliability exceeded 0.7, indicating acceptable consistency, and a second round of coding was performed for the final analysis. Table IV presents the results of the coding analysis.

We observed different patterns of question prompt usage across the three prompt-enabled UI conditions. In the SCAFFOLDING condition, four of the five participants described a "question-driven search" strategy, actively using the displayed questions to guide their web searches. For example:

- (P1-Scaffolding) *I asked myself the prompt questions and searched for information to support my answers.*
- (P24-Scaffolding) *When performing additional searches, I input search terms inspired by the questions that appeared.*

A few participants under the ADJUNCT and EXPECTED conditions also reported engaging in question-driven search, but the frequency was lower (ADJUNCT: 1/5; EXPECTED: 1/7).

Under the EXPECTED condition, five of the seven participants reported that the question prompts made them "aware of unexpected perspectives." Representative comments include:

- (P18-Expected) *The prompts included questions I would not have thought of myself, like how to connect economic support with educational needs. They helped me deepen my opinion.*
- (P19-Expected) *When I saw the prompt about effective volunteer-based learning support, I incorporated the idea, which I hadn't previously considered.*

These participants reported that the prompts helped them notice new viewpoints and integrate them into their information seeking and opinion development. These results indicate that the EXPECTED prompts were useful for inquiry-oriented search.

In contrast, only a few participants under the ADJUNCT (2/5) and SCAFFOLDING (1/5) conditions were aware of unexpected perspectives. Some participants noted that although they noticed new viewpoints, they did not actively incorporate them into their opinions. For example:

- (P5-Adjunct) *I looked at the prompts as mere references. Still, I thought the questions could lead to valuable discussions.*

Under the SCAFFOLDING and ADJUNCT conditions, the prompts served as "triggers for deeper exploration" of the sought topics for several participants (SCAFFOLDING: 2/5; ADJUNCT: 2/5). However, the cognitive process leading to this outcome differs among the UI conditions. Under the SCAFFOLDING condition, participants reported that the prompts directly motivated them to explore related topics. In contrast, participants under the ADJUNCT condition described revisiting or recalling recently viewed webpages in response to the prompts, which led to deeper exploration. For example:

- (P1-Scaffolding) *I asked myself the prompt questions and searched for information to support my answers.*
- (P14-Adjunct) *I often returned to webpages to verify the prompt content. For instance, the question "How do families on welfare make inquiries?" made me focus on the topic of parental applications and inspired further exploration.*

Some participants under the SCAFFOLDING and EXPECTED conditions also noted that the prompts were useful in "organizing their opinions" for the task (SCAFFOLDING: 2/5; EXPECTED: 2/7). Below are representative comments:

- (P12-Scaffolding) *I thought the prompts might reflect the final questions I'd need to answer, so I prepared accordingly and eventually synthesized opinions for some of them.*
- (P17-Expected) *When summarizing my opinion, I referred to the prompts. I think the presence of these questions helped me create a cohesive summary.*

Only participants using the ADJUNCT UI reported that prompts served as "cues for recalling previously read information" (2/5):

- (P14-Adjunct) *The question "What are the main support activities of Kids' Door?" reminded me of something I*

had read earlier. The question made me want to revisit and confirm the details.

- (P16-Adjunct) *I often went back to webpages to check the content related to the prompt.*

As described above, the participants demonstrated diverse usage patterns of prompting questions. Some participants actively leveraged the prompts, whereas others viewed them as mere references. Negative feedback was also reported, particularly under the ADJUNCT and EXPECTED conditions. Participants noted that the prompts were sometimes irrelevant or uninteresting (ADJUNCT: 3/5; EXPECTED: 2/7). In addition, under the EXPECTED condition, one participant reported being confused when the question prompts conflicted with their existing opinion:

- (P2-Expected) *When a prompt contradicted my existing idea, I was not sure if I should revise my opinion or follow the suggestion. It made me spend a lot of time thinking.*

VI. DISCUSSION

For **RQ1**, the results of the online study indicate that participants under the SCAFFOLDING condition spent more time per query on the SERP than those under the PLAIN condition. A similar tendency was observed for the ADJUNCT condition. In contrast, although no significant difference was observed in the SERP duration per query under the EXPECTED condition, the total SERP duration was longer than that under the PLAIN condition. These results suggest that under the SCAFFOLDING, EXPECTED, and ADJUNCT conditions, participants may have spent more time on the SERP screen because they reflected on the asked questions or scrutinized the search results prompted by the displayed questions.

For **RQ2**, the online study revealed that the question prompts under the SCAFFOLDING and EXPECTED conditions were perceived as more relevant and important to the inquiry theme than those under the ADJUNCT condition. In addition, the prompts under the SCAFFOLDING condition were evaluated as more interesting and useful in gaining new perspectives on the theme than those under the ADJUNCT condition. Furthermore, the prompts under the EXPECTED condition were rated more helpful than those under the ADJUNCT condition to gain new perspectives and organize the collected information and personal opinions. However, the results of the laboratory study demonstrate that the way participants used the question prompts varied based on the UI condition.

For **RQ3** related to question prompt usage, the prompts under the SCAFFOLDING and EXPECTED conditions helped users organize the collected information and their opinions. Analysis of the relevance between the displayed prompts and the opinions reported by the participants revealed that using the SCAFFOLDING and EXPECTED conditions may promote the tendency to formulate opinions related to the prompts displayed during the search task compared with the ADJUNCT condition. However, although the EXPECTED condition directed user attention to perspectives they had not previously considered, the SCAFFOLDING condition supported deeper exploration of already-considered perspectives, thereby encouraging more active information-seeking and reflection during the search. These differences can be attributed to the pedagogical nature of the EXPECTED condition, in which the system defines a comprehensive set of questions based on elements deemed relevant to the inquiry theme, compared with the constructivist approach [13] of the SCAFFOLDING condition, which provides context-aware question prompts to stimulate deeper thinking.

For the EXPECTED condition, the lab study indicated that they helped broaden users' perspectives by drawing their attention to unexpected viewpoints; however, there were concerns that irrelevant or conflicting prompts could confuse the participants. The EXPECTED condition displayed prompts randomly selected from a pregenerated pool, whereas the SCAFFOLDING and EXPECTED conditions displayed prompts each time users viewed a webpage. Therefore, to more effectively support inquiry-oriented search, it is necessary to place the SCAFFOLDING condition, which facilitates active search and reflection, at the core of the design while considering (1) combining the context-aware prompts of the SCAFFOLDING condition, which promote deeper inquiry, with the perspective-broadening question prompts of the EXPECTED condition and (2) optimizing the timing of the prompt display.

In the online study, the participants perceived the prompts displayed in the ADJUNCT condition as less relevant or important to the inquiry theme. Similarly, in the lab study, a few participants actively used the prompts during their inquiry. The lab study also revealed that the ADJUNCT prompts help users recall information they had already read. Because the question prompts are generated in reverse from the previously viewed pages, they tend to prompt users to recall or reconfirm the page content. Although this may not directly contribute to inquiry-oriented search tasks aimed at opinion formation through the analysis and synthesis of information, it can help users understand and memorize fundamental knowledge, which serves as a basis for deeper inquiry.

Implications for digital libraries

Deployed in web search engines and digital library IR tools, our context-aware scaffolding prompts can strengthen information literacy by nudging searchers to interrogate sources, surface counter-arguments, or connect cross-disciplinary concepts during search. Beyond fact lookup, such scaffolding supports higher-order inquiry skills and may help mitigate confirmation bias in everyday information seeking.

Limitations and future work

The approach proposed in this study and the experimental methodology have several limitations. The first limitation is the identification of users' information-seeking modes. Here, we assumed inquiry-oriented search tasks in the experiments; however, in real-world web search scenarios, users do not always engage in search tasks. In some cases, lookup or exploratory search modes may be sufficient, and users may switch between different search modes, including inquiry-oriented search. Therefore, it is necessary to identify users'

information-seeking modes during a web search to appropriately apply the proposed prompting strategies.

The second limitation involves analyzing how users examine and organize the collected information and form their opinions. The behavioral analysis focused only on activities within the experimental website and did not consider users' interactions on external webpages reached through the search results. To compensate for this, in the laboratory study, participants were asked to retrospectively describe their search behavior and use of question prompts during the tasks. However, to more rigorously analyze the processes of information seeking and opinion formation, think-aloud protocols should be employed, and participants should be shown recordings of their task performances and their reports.

The third limitation involves prompt generation. To address the challenge of not being able to directly observe users' thoughts during a web search, the proposed method employs an LLM to predict opinions that a hypothetical learner might form when reading a page and then generates prompts accordingly. However, because the opinions users form may vary depending on their demographic attributes and prior knowledge, the relevance and effectiveness of the prompts may also differ among users. Thus, the personalization of prompt generation should be considered in future studies.

The unidirectional nature of the interaction is also a challenge. The proposed system presents question prompts to users but does not process their direct responses to the prompts. Some users may expect to receive system responses after answering a prompt. Users' responses to the prompts can contribute to a better understanding of their thought processes and help improve or personalize prompt generation.

VII. Conclusion

Herein, we propose a method that generates context-aware scaffolding prompts to facilitate inquiry-oriented search. The study results demonstrate that presenting scaffolding prompts can increase the time users spend on the search engine result page per query, promote question-driven information seeking, and encourage behavior that deepens the understanding of the task theme under exploration. In addition, when prompts are generated by decomposing the elements necessary for thematic exploration without considering users' search contexts, users' attention is drawn to perspectives they have not considered. In future studies, we will combine scaffolding and pedagogical prompting to optimize the timing of prompt display.

Acknowledgment

The work was supported by the Grants-in-Aid for Scientific Research (23K21725, 25K03228, 25K03229) from the MEXT of Japan and JST CREST (JPMJCR2562).

References

[1] Y. Benjamini and Y. Hochberg, "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 1, pp. 60–83, 2000.

[2] A. Câmara, N. Roy, D. Maxwell, and C. Hauff, "Searching to learn with instructional scaffolding," in *CHIIR*, 2021, pp. 209–218.

[3] Z. Cheng, J. Xu, and H. Jin, "Treequestion: Assessing conceptual learning outcomes with llm-generated multiple-choice questions," in *CSCW*, 2024, pp. 1–29.

[4] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, and R. Syed, "Assessing learning outcomes in web search: A comparison of tasks and query strategies," in *CHIIR*, 2016, pp. 163–172.

[5] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes, "Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations," in *CHI*, 2023, pp. 1–13.

[6] Y. Gu, "To code or not to code: Dilemmas in analysing think-aloud protocols in learning strategies research," *System*, vol. 43, pp. 74–81, 2014.

[7] M. Harvey, C. Hauff, and D. Elsweiler, "Learning by example: Training users with high-quality query suggestions," in *SIGIR*, 2015, pp. 133–142.

[8] J. Krajcik, K. L. McNeill, and B. J. Reiser, "Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy," *Science Education*, vol. 92, no. 1, pp. 1–32, 2008.

[9] Y. Lee, T. S. Kim, S. Kim, Y. Yun, and J. Kim, "Dapie: Interactive step-by-step explanatory dialogues to answer children's why and how questions," in *CHI*, 2023, pp. 1–22.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.

[11] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *ACL*, 2020, pp. 1906–1919.

[12] F. Moraes, S. R. Putra, and C. Hauff, "Contrasting search as a learning activity with instructor-designed learning," in *CIKM*, 2018, pp. 167–176.

[13] J. Piaget, M. Cook *et al.*, *The origins of intelligence in children*. International universities press New York, 1952, vol. 8, no. 5.

[14] B. J. Reiser, "Scaffolding complex learning: The mechanisms of structuring and problematizing student work," *Journal of the Learning Sciences*, vol. 13, no. 3, pp. 273–304, 2004.

[15] S. Y. Rieh, K. Collins-Thompson, P. Hansen, and H.-J. Lee, "Towards searching as a learning process: A review of current perspectives and future directions," *Journal of Information Science*, vol. 42, no. 1, pp. 19–34, 2016.

[16] C. Rosset, C. Xiong, X. Song, D. Campos, N. Craswell, S. Tiwary, and P. Bennett, "Leading conversational search by suggesting useful questions," in *WWW*, 2020, pp. 1160–1170.

[17] N. Sharma, Q. V. Liao, and Z. Xiao, "Generative echo chamber? effect of llm-powered search systems on diverse information seeking," in *CHI*, 2024, pp. 1–17.

[18] M. Suzuki and Y. Yamamoto, "Characterizing the influence of confirmation bias on web search behavior," *Frontiers in Psychology*, vol. 12, pp. 1–11, 2021.

[19] K. Umemoto, T. Yamamoto, and K. Tanaka, "Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search," in *SIGIR*, 2016, pp. 405–414.

[20] R. White, "Beliefs and biases in web search," in *SIGIR*, 2013, pp. 3–12.

[21] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," *Journal of child psychology and psychiatry*, vol. 17, no. 2, pp. 89–100, 1976.

[22] Y. Yamamoto, "Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making," *Frontiers in Psychology*, vol. 15, pp. 1–16, 2024.

[23] Y. Yamamoto and T. Yamamoto, "Query priming for promoting critical thinking in web search," in *CHIIR*, 2018, pp. 12–21.

[24] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, and S. Dietze, "Predicting user knowledge gain in informational search sessions," in *SIGIR*, 2018, pp. 75–84.

[25] C. Zhang, X. Liu, K. Ziska, S. Jeon, C.-L. Yu, and Y. Xu, "Mathemyths: Leveraging large language models to teach mathematical language through child-ai co-creative storytelling," in *CHI*, 2024, pp. 1–23.

[26] S. Zhao, H. Wang, C. Li, T. Liu, and Y. Guan, "Automatically generating questions from queries for community-based question answering," in *IJCNLP*, 2011, pp. 929–937.

[27] P. Zhu, A. Câmara, N. Roy, D. Maxwell, and C. Hauff, "On the effects of automatically generated adjunct questions for search as learning," in *CHIIR*, 2024, pp. 266–277.