On the Effects of Automatically Generated Adjunct Questions for Search as Learning

Zhu, Peide; Câmara, Arthur; Roy, Nirmal; Maxwell, David; Hauff, Claudia

DOI
10.1145/3627508.3638332

Publication date
2024

Document Version
Final published version

Published in
CHIIR 2024 - Proceedings of the 2024 Conference on Human Information Interaction and Retrieval

Citation (APA)
Zhu, P., Câmara, A., Roy, N., Maxwell, D., & Hauff, C. (2024). On the Effects of Automatically Generated Adjunct Questions for Search as Learning. In *CHIIR 2024 - Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (pp. 266-277). (CHIIR 2024 - Proceedings of the 2024 Conference on Human Information Interaction and Retrieval). ACM. https://doi.org/10.1145/3627508.3638332

On the Effects of Automatically Generated Adjunct Questions for Search as Learning

Peide Zhu*
p.zhu-1@tudelft.nl
Delft University of Technology
Delft, The Netherlands

Arthur Câmara†
camara@zeta-alpha.com
Zeta Alpha
Amsterdam, The Netherlands

Nirmal Roy
n.roy@tudelft.nl
Delft University of Technology
Delft, The Netherlands

David Maxwell†
maxwelld90@acm.org
Booking.com
Amsterdam, The Netherlands

Claudia Hauff†
claudia.hauff@gmail.com
Spotify
Delft, The Netherlands

ABSTRACT

Actively engaging learners with learning materials has been shown to be very important in the Search as Learning (SAL) setting. One active reading strategy relies on asking so-called *adjunct questions*, i.e., manually curated questions geared towards essential concepts of the target material. However, manual question creation is impractical given the vast online content. Recent research has explored the effects of Automatic Question Generation (AQG) on aiding human learning. These studies have primarily focused on user studies in controlled online reading scenarios with limited documents. However, the impacts of adjunct questions on learning in the SAL setting, which involves learning through web searching, are not yet well understood. This paper addresses this gap by conducting a user study with automatically generated adjunct questions integrated into the reading interface built on top of a search system. We conducted a between-subjects user study ($N = 144$) to investigate the incorporation of automatically generated adjunct questions on participants' learning. We employed three different question generation strategies as well as a control condition: (i) synthesis questions; (ii) factoid questions targeting random text spans; and (iii) factoid questions targeting terms and phrases relevant to the information need at hand. We present four major findings: (i) participants who received adjunct questions exhibited significantly more fine-grained reading behaviour, such as longer document dwell time and more scrolls, than those without adjunct questions. However, adjunct questions' influence on learning outcomes depends on the AQG strategy. (ii) Question types significantly influence participants' reading behaviour. (iii) The adjunct questions' target spans significantly influence learning outcomes. Lastly, (iv) participants' prior knowledge levels affect adjunct questions' effects on their learning outcomes and their reaction to different AQG strategies. Our findings have significant design implications for learning-oriented search systems. The data and code is available at https://github.com/zpeide/AQG-AdjunctQuestions.

ACM Reference Format:
Peide Zhu, Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2024. On the Effects of Automatically Generated Adjunct Questions for Search as Learning. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24), March 10–14, 2024, Sheffield, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627508.3638332

1 INTRODUCTION

Searching and reading online materials has become a crucial way of learning. It is generally considered inefficient, though, to learn by passively browsing and reading documents [24, 58]. In contrast, actively engaging learners during this process with retrieval practice methods like *adjunct questions*— i.e., asking questions about specific parts of a document to draw attention to the reading materials [24] and retrieving information from one's memory [12, 37]—leads to better learning outcomes. Extensive research on the effects of asking such questions has been conducted, and generally, it has been found to have a positive effect on learning. However, these studies have mostly been conducted in controlled classroom settings [35, 56, 57] and with *manually* curated questions.

Considering the amount of content available on the web, this is not a feasible approach in the Search as Learning (SAL) setting, where learning behaviour commonly involves searching over open-domain resources, targeting complex concepts instead of fact-finding, and learning by reading and integrating knowledge across documents [11, 16, 47, 71]. With the ever-improving generation quality of pre-trained language models (PLM), some works have analyzed the effectiveness and potential benefits of Automatic Question Generation (AQG) on human learning. For example, Syed et al. [68], Van Campenhout et al. [70] demonstrated that automatically generated questions performed comparably to human-authored questions. Moreover, some works [66, 68] highlighted the potential importance of incorporating automatically generated adjunct questions. Notably, Syed et al. [68] found that in the context of reading comprehension, learners who received automatically generated adjunct questions spent more time reading and paid more attention to the reading materials than those without such questions. Additionally, the impact of adjunct questions on learning outcomes varied

*The first author has been supported by the China Scholarships Council (CSC).

†This work is not related to the activities of the current affiliations (Zeta Alpha, Booking.com, and Spotify).

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

ACM ISBN 979-8-4007-0434-5/24/03
https://doi.org/10.1145/3627508.3638332

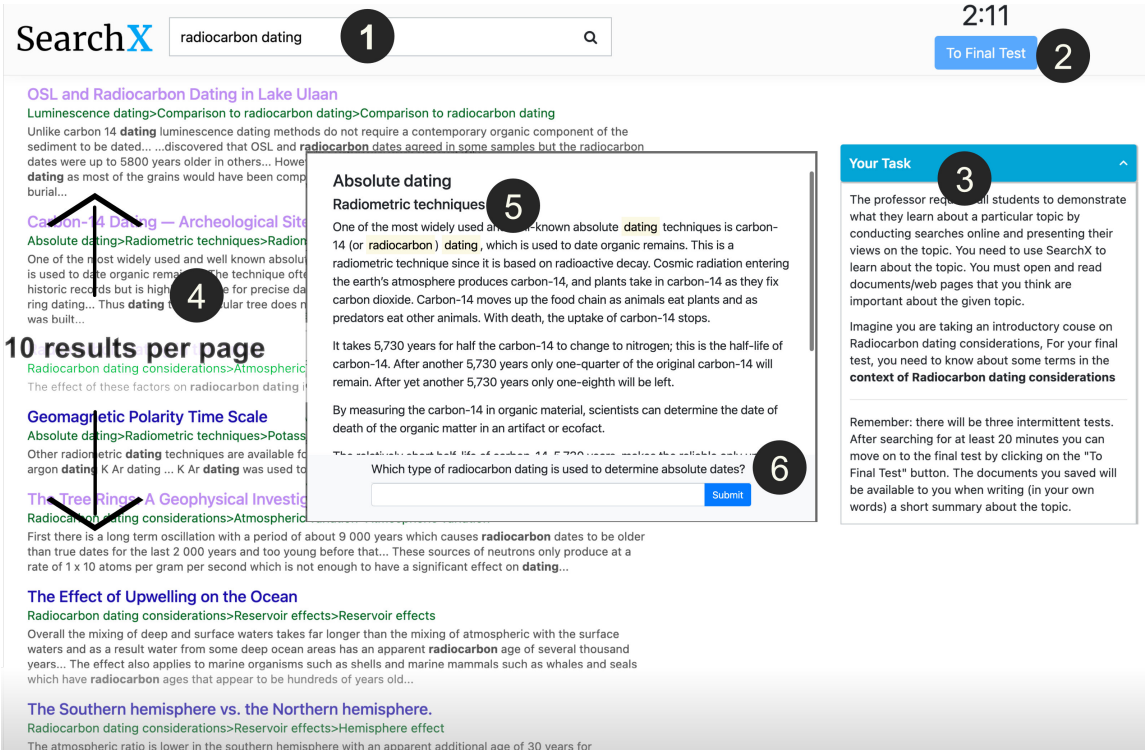


Figure 1: Screenshot of the system interface used by participants for searching and learning on the assigned topic (e.g., *radiocarbon dating considerations*). The circled numbers correspond to the narrative of Section 3.1.

depending on learners' prior knowledge. Learners with low prior knowledge benefited from adjunct questions significantly in terms of long-term retention, in contrast to those with human-curated or without adjunct questions.

Despite these insights, two major limitations remain. First of all, as a crucial way of learning, the searching component is absent from these studies [66, 68, 70]: they focused on the controlled online reading scenario with a relatively small number (around 100) of questions. Secondly, these studies evaluate learning outcomes with factoid questions instead of higher-level skills like writing. Learning through web search engines significantly differs from the controlled reading comprehension settings—learners choose their own queries, have access to a much larger body of documents, and self-select documents to view—the effect of including adjunct questions in the SAL setting is not yet well understood.

In our study, we address this gap by investigating the effects and factors of actively involving learners with adjunct questions integrated within the reading interface built on top of a search system. We aim to assess the impact of this setup on both learners' behaviour and learning outcomes in the context of learning-oriented search tasks. We implemented a search system that supports adjunct questions with a corresponding UI widget on top of the open-source SearchX system [52], as shown in Figure 1. We conducted a between-subjects user study with $N = 144$ participants, where participants were assigned to one of four variants: *(i)* $\mathbf{Q}_{none}$*, the control condition without adjunct questions; (ii)* $\mathbf{Q}_{synthesis}$*, synthesis questions; (iii)* $\mathbf{Q}_{random}$*, factoid questions targeting random text spans; and (iv)* $\mathbf{Q}_{term}$*, factoid questions targeting terms and phrases relevant to the information need at hand.* Participants' learning outcomes were measured through two tasks: a recall-based vocabulary learning task [11, 60, 73], and an essay writing task [11, 44, 59] that involved higher cognitive complexity. With this user study, we aim to answer the following research questions *in a SAL setting*:

RQ1 *To what extent do automatically generated adjunct questions impact participants' behaviour and learning outcomes?*
RQ2 *How do characteristics of adjunct questions and participants' prior knowledge affect participants' learning?*

We study two important question characteristics, including the question types (factoid vs. synthesis) and the question target selection. Factoid questions require only the extraction of basic facts. Synthesis questions require higher-level cognitive skills like integrating, evaluating, and analyzing different facts.

Overall, we present four major findings. **(i)** Compared to the control condition, adjunct questions have a significant influence on participants' behaviour, with participants in these settings displaying more fine-grained reading, as evidenced by longer reading times and scrolls, as well as fewer queries in the search session. **(ii)** The question types have a significant influence on participants' reading behaviour. With synthesis questions, participants achieve better learning outcomes on the task that requires higher cognitive complexity compared to those required to answer factoid questions regarding random text spans. **(iii)** The target spans of adjunct questions (random vs. focused) have a significant influence on learning

outcomes. **(iv)** Participants' prior knowledge levels affect adjunct questions' effects on their learning outcomes and reactions to different AQG strategies. Participants with higher prior knowledge, in general, achieve better learning gains.

These findings provide empirical evidence that in order to incorporate adjunct questions into a learning-oriented searching system effectively, it is essential to identify learners' learning targets and their prior knowledge and generate types of questions accordingly.

2 BACKGROUND

2.1 Search as Learning

Unlike traditional ad-hoc search systems that generally consider a user's information need as atomic (i.e., a single information need is covered by a single user query) [5, 6, 16], a search system designed for SAL must be aware of the nature of users' tasks [10, 47, 50, 71], as these may encompass multiple rounds of interaction with the system, with varying degrees of complexity. Over the past decade, SAL has attracted considerable attention, and many different approaches which touch different parts of the search system to help users learn while searching have been proposed.

Backend adaptations. Search systems are naturally complex, with multiple components working together to help the user search for relevant documents. While many prior SAL works have focused on front-end adaptations (as we will discuss below), studies investigating how changes made directly to the retrieval pipeline impact learner's behaviour are still rare. For example, Syed and Collins-Thompson [67] designed a retrieval algorithm to improve the ranking of documents with a higher density of vocabulary terms related to the learner's topic. Collins-Thompson et al. [15] demonstrated how tweaking the ranking system according to the learner's reading level can also be beneficial. Finally, Athukorala et al. [3] showed that a reinforcement-learning-based ranking algorithm can improve the learner's experience by balancing the diversity or depth of the search results according to the learner's intention.

Frontend adaptations. Most prior SAL works have focused on aiding users in writing queries and organizing thoughts and content. Learning-oriented adaptations to the Search Engine Results Page (SERP), such as displaying an outline of the topic the learner is interested in [11], providing entity cards [61], or including conversational interfaces [54] have been shown to help users with their knowledge acquisition process—at least to some extent. Approaches that help formulate queries have also been studied, as learners' querying behaviour plays a vital role in their learning process [43, 74]. For instance, inspired by [69], Câmara et al. [11] displayed a progress bar that estimates how much topic exploration has been done, considerably influencing learners' querying behaviour. Another type of change made to the UI is related to how learners organize their materials and thoughts, as explored by [60] where learners were prompted to highlight parts of the text they may find relevant and take notes directly on the SERP. Liu et al. [45] asked learners to build mind-maps, leading to a measurable change in their behaviour and knowledge gains.

Active learning. Some of the strategies mentioned above are examples of educational active learning techniques. Instead of passively reading, the learner *actively* engages with the learning material. These strategies have consistently been shown to considerably improve learner's knowledge retention [26, 29, 62]. A popular method of implementing active learning is asking learners questions about the material they come across during the search process. These questions are designed to guide learners' attention to specific portions of the material (ideally those covering the key ideas) and, therefore, help learners to understand and remember the material better [38, 58]. The effects of using questions to foster learning (i.e., adjunct question effects) are well known in the classroom setting [2, 35, 56, 57]. Importantly, these questions are typically created manually by topical and educational experts. This is an expensive and slow process and not feasible to do at scale, considering the quantity and diversity of online learning materials. In contrast, automatic question generation is scalable.

2.2 Automatic Question Generation

As a critical Natural Language Processing (NLP) task, AQG has been heavily researched over the past decades. Various template-based [27, 36, 49] and neural network-based [9, 23, 25] methods have been proposed. Like other NLP tasks, with the advance of PLM, AQG approaches have jumped considerably in quality as measured by automatic metrics and human evaluations [4, 20, 22, 41, 42, 53]. Some works have investigated the application of AQG to education [1, 8, 13, 19, 39, 40, 64, 65, 72]. These prior works though, focus mainly on how to apply AQG methods to educational materials and how to generate various types of questions for educational purposes. *The effects of automatically generated questions on human learning still need to be well investigated.*

Several works [33, 46, 66, 68] have recently begun to study this question. In particular, Syed et al. [68] systematically analyzed the effectiveness of AQG on human learning compared to manually curated questions, as well as other impact factors such as learners' prior knowledge, the type of adjunct questions (factoid or synthesis), and the content that questions focused on. Like Syed et al. [68], Steuer et al. [66] studied automatically generated adjunct questions' effects on non-native speakers' English vocabulary learning. The effects were evaluated by the self-report of prior knowledge on the topic and the correctness of post-test questions. Van Campenhout et al. [70] used automatically generated questions in a university course as formative practice and evaluated the questions' effects by measuring the students' behaviour such as engagement in practice. However, these works were conducted in a *controlled scenario* by showing participants *one Wikipedia article* or *a fixed list of documents* and corresponding questions (around 100).

Finally, we point out that in our work, we considered two types of questions: factoid questions and synthesis questions. Factoid questions focus on specific facts in the document; these questions primarily address the *Remembering* level of cognitive complexity in Bloom's taxonomy [7]. In contrast, synthesis questions require higher levels of cognitive complexity like *Analyze* and *Evaluate*. In Syed et al. [68], while the factoid questions were automatically generated, the synthesis questions were not. In our work, we extend prior work in two directions: (i) we instantiate the concept of adjunct questions in an actual search system, and (ii) we automatically

generate different types of questions and investigate their effect on behaviour & learning.

3 ADJUNCT QUESTIONS IN SEARCHX

3.1 SearchX Interface

To carry out this study, inspired by [11, 60, 68], we used SearchX [52], a modular, open-source framework that supports IR experiments. SearchX contains a number of modern search engine front-end features and widgets akin to a contemporary web search engine's SERP. Moreover, combined with LogUI [48], it offers fine-grained search logs (hovers, clicks, scrolls, etc.). Figure 1 shows the interface we implemented for our experiments. ❶ represents the query box (without query auto-completion). ❷ denotes the timer to help participants count the task time. After the search session lasts at least 20 minutes, the *To Final Test* button becomes available and leads the participant to the post-test when clicked. The task description is shown in ❸, where the assigned topic is bold-faced. ❹ represents the search results page. We show 10 results per page and up to 5 pages, which we consider sufficient search depth as participants only sometimes go beyond the second page [34, 50]. The search results are provided by ElasticSearch[1]. Notably, we show a short snippet created by extracting document sentences containing content words of the query in order to provide participants with essential information. Once the participant clicks on a link, a scrollable document viewer ❺ pops up and displays the document. At the bottom of the viewer is the AQG widget ❻, which is invisible for participants in the control condition ($\mathbf{Q}_{\text{none}}$). In the other three conditions, it shows one automatically generated question about the document. A participant can only proceed to another document or the SERP if they provide *some* answer to the question. The answer correctness does not affect participants' payment and is only used for further analysis.

3.2 Automatic Adjunct Question Generation

3.2.1 Dataset. Realistic learning by searching involves searching, reading, and gathering knowledge over large-scale open-domain documents. While we could have opted for a web search API as a retrieval backend, this was not feasible as we could not generate questions at scale from any website within a few milliseconds. Instead, we selected a corpus and pre-computed the questions of each type. Specifically, we used the `benchmarkY1train` set of the TREC-CAR v1.5 dataset [21]. This dataset contains a set of structured Wikipedia topics with headings designed to retrieve answers for complex information needs—it has been used in prior SAL research [11, 60]. We used 117 topics in the `benchmarkY1train` set and the 91 vocabulary terms (representative concept phrases) created by Câmara et al. [11]. Additionally, we extracted 136 topics from the TREC-CAR `train-v1.5` set that contained the vocabulary terms to ensure participants were exposed to plenty of documents containing the target vocabulary terms. In total, we used 253 Wikipedia topics. As each topic corresponds to one long Wikipedia article that requires considerable reading time, as shown in [68], we split articles into 1,627 documents based on their heading structures to engage participants with more searching and reading behaviour.

We studied two categories of questions and employed separate question generators for each: (i) factoid (or low-level) questions and (ii) synthesis (or high-level) questions. As illustrated in Table 1, factoid questions seek text spans that pertain to specific facts, such as concepts and numbers, which can be directly retrieved from the text. In contrast, synthesis questions necessitate comprehensive efforts, such as integrating and analyzing document information, surpassing the mere extraction of text spans.

3.2.2 Factoid Question Generation. We used the PAQ [42] framework for generating factoid questions. First, we utilized two extraction methods provided by PAQ to identify text spans within a document that are worth questioning. One method involved extracting all named entities as potential answers, as named entities such as names, numbers, and locations often convey significant information. The other method involved a trained neural model as the answer span extractor called `Span2DAnswerExtractor`[2]. In addition, we also included all vocabulary terms as question-worthy text spans. We opted for the `qgen_multi_base`[3], a BART [41]-based model fine-tuned on various QA datasets as the factoid question generator. It took the document and extracted text spans as inputs, resulting in 65,237 questions for the 1,627 documents along with the 253 topics. The generated questions underwent a filtering process concerning question length and consistency. First, questions shorter than 6 words were disregarded, resulting in the removal of 392 questions. Subsequently, the remaining questions were filtered using PAQ's QA-Pair filtering tool, which assessed the consistency between the answer and the generated question. This step led to the further filtering of 37,016 questions. If multiple valid QA pairs existed for a single document, we selected the pair with the highest answer score for that document. We then separated all factoid questions into two groups: questions regarding the vocabulary terms ($\mathbf{Q}_{\text{term}}$, 750, covering 64 terms) and other text spans ($\mathbf{Q}_{\text{random}}$, 1,627). Although answers for $\mathbf{Q}_{\text{random}}$ tend to be informative and important, they were extracted regardless of the participants' learning goals.

3.2.3 Synthesis Question Generation. Synthesis questions typically require more than text spans from the documents to provide comprehensive answers. PAQ is primarily trained to cater to factoid questions, so it may not be well-suited for generating synthesis questions. To address this limitation, we opted to fine-tune the BART model [41] using the ELI5 dataset [28] that comprises complex, diverse questions that require long-form multi-sentence answers, e.g., *Why are flutes classified as woodwinds when most of them are made out of metal?*, aligning with the requirements of synthesis question generation. We generated one synthesis question for each document paragraph, resulting in 5,393 synthesis questions. Among the questions of the same document, we selected the longest one as the synthesis question for the study, i.e., we kept 1,627 synthesis questions.

Table 1 shows examples of our generated factoid and synthesis questions. As shown in these examples, facts to answer the $\mathbf{Q}_{\text{random}}$

[1] https://www.elastic.co/

[2] https://github.com/facebookresearch/PAQ#answer-extraction

[3] https://github.com/facebookresearch/PAQ

Table 1: An example of automatically generated questions from a given document. Shown here are two factoid questions (Q_{random}, Q_{term}) and a synthesis question ($Q_{synthesis}$). Highlighted in cyan and pink are answers for creating the corresponding factoid questions. Extracted word spans that are filtered out are highlighted in violet.

Example	Irritable bowel syndrome	
Document	Approximately 10 percent of IBS cases are triggered by an acute gastroenteritis infection. Genetic defects relating to the innate immune system and epithelial barrier as well as high stress and anxiety levels appear from evidence to increase the risk of developing post-infectious IBS. Post-infectious IBS usually manifests itself as the diarrhea predominant subtype. Evidence has demonstrated that the release of high levels of proinflammatory cytokines during acute enteric infection causes increased gut permeability leading to translocation of the commensal bacteria across the epithelial barrier resulting in significant damage to local tissues which is likely to result in chronic gut abnormalities in sensitive individuals. However, increased gut permeability is strongly associated with IBS regardless of whether IBS was initiated by an infection or not.	
Q_{random}	What percentage of ibs cases are triggered by an acute gastroenteritis infection?	10 percent
Q_{term}	What part of the gut is affected by irritable bowel syndrome?	epithelial barrier
$Q_{synthesis}$	Why do some people develop IBS more often than others?	

and Q_{term} questions can be directly found in the document as text spans. In contrast, the generated synthesis questions require comparing and analyzing document contents.

3.3 Question Quality Evaluation

To ensure the quality of the generated questions, we randomly sampled 30 generated questions from Q_{random}, Q_{term}, and $Q_{synthesis}$, respectively, in addition to 30 human-curated questions from the 4 SQuAD [55] articles used in [68] for comparison. We conducted a human evaluation by recruiting five native English speakers with at least undergraduate degrees as annotators. The questions were rated on a 5-point scale concerning their relevance to their context, the answerability, i.e., whether they can be answered with information from the document, and the possibility that a human wrote the question. The final rating of each question is determined by averaging all annotators' ratings. Table 2 reports the average score along all these measures. We conducted a one-way ANOVA test on the measures with respect to the question type factor. First, the average length of questions ranged from 11.3 to 13.2, and there was no significant difference. Second, we can observe that although automatically generated questions were considered less human-written, they were still considered as likely written by humans (> 3.4 on a 5-point scale, compared to 4.28 for human-curated SQuAD questions). Furthermore, $Q_{synthesis}$ questions were significantly lower than the SQuAD questions in terms of relevance ($p < 10^{-4}$) and answerability ($p < 10^{-4}$). One possible reason is synthesis questions tend to require more cognitive complexity and background knowledge than SQuAD questions which are simple factoid questions. Notably, the answerability of Q_{random} questions was significantly higher than that of synthesis questions ($p = 0.034$). This is aligned with our design since the synthesis questions are supposed to be more challenging to answer.

Table 2: Comparison of SQuAD and automatically generated questions in terms of avg. question length and human evaluation for Relevance, Answerability, and Human-Written (H-W) on a 5-point scale. † denotes the one-way ANOVA significance, while $\mathcal{U}$(SQuAD), $\mathcal{S}$($Q_{synthesis}$), $\mathcal{R}$(Q_{random}), $\mathcal{T}$(Q_{term}) indicate post-hoc significance (TukeyHSD pairwise test, p<0.05) over four groups of questions.

Method	**Length**	**Relevance**†	**Answerability**†	**H-W**†
SQuAD	11.3	$4.75^{\mathcal{TS}}$	$4.68^{\mathcal{TS}}$	$4.28^{\mathcal{TS}}$
Q_{random}	13.2	4.34	$4.10^{\mathcal{S}}$	3.79
Q_{term}	12.8	$4.03^{\mathcal{U}}$	$3.55^{\mathcal{U}}$	$3.55^{\mathcal{U}}$
$Q_{synthesis}$	12.0	$3.87^{\mathcal{U}}$	$3.47^{\mathcal{UR}}$	$3.44^{\mathcal{U}}$

4 USER STUDY DESIGN

4.1 Topics

In line with prior research [11, 44, 60], we designed two learning-focused tasks to assess participants' learning outcomes: a recall-based vocabulary learning task and an essay writing task. The vocabulary-learning task assessed knowledge levels on vocabulary terms at cognitive levels like remembering and understanding based on revised BLOOM's taxonomy [7]. On the other hand, the essay writing task required participants to compose a summary of at least 100 words based on their acquired knowledge during the search session. This task aimed to assess higher cognitive levels, such as evaluating and analyzing. We chose seven topics from the 117 topics in `benchmarkY1train` along with their vocabulary terms for the learning tasks. These topics have suitable complexity, so they are not too easy that most participants already have plenty of knowledge or are too hard to learn in twenty minutes. Table 3 presents the topics and vocabulary terms.

4.2 Experimental Conditions

As mentioned earlier, in our user study, we assign each participant to one of four conditions:

Q_{none} In the control condition, we do not show participants the AQG widget (❻ in Figure 1) in the document viewer.

$Q_{synthesis}$ In this condition, we present a participant with a high-level synthesis question about the opened document.

Q_{random} In this condition, for each document a participant opens, we present one automatically generated factoid question regarding a text span like one named entity randomly sampled from the document.

Q_{term} In this condition, if there are vocabulary terms of the assigned topic in the document, we present the participant with an automatically generated factoid question regarding one of the vocabulary terms. Otherwise, a random factoid question would be presented instead.

4.3 Study Workflow

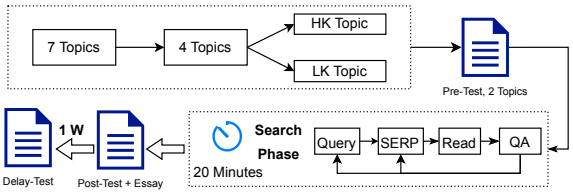


Figure 2: Illustration of the user study workflow. This flow describes the experimental conditions of Q_{random}, $Q_{synthesis}$, and Q_{term}. The Q_{none} condition does not take the QA step.

We now briefly introduce the experimental procedure that consists of seven phases.

1. **Task Introduction.** Participants read a general introduction to the entire study workflow.
2. **Survey.** Participants were asked to complete a demographics survey containing questions regarding their education level, language skills, and their use of web search engines and online documents for learning.
3. **Topic Selection.** We selected seven topics for our user study. To prevent familiarity bias [32], we designed a two-step knowledge selection procedure. First, we randomly chose four of seven topics and asked participants to choose one topic they knew best and one they knew least. Both of the topics were used for the vocabulary knowledge pre-test.
4. **Pre-Test.** Participants completed two vocabulary knowledge tests. Each test consisted of 10 vocabulary questions on the selected topics. We randomly chose one topic with equal probability as the assigned topic to learn more about in the search phase.
5. **Search Phase** We randomly assigned each participant to one of our four conditions. Participants needed to spend at least 20 minutes searching and reading documents to learn about the assigned topic in line with prior works [50, 68].
6. **Post-test.** After 20 minutes in the search phase, participants could continue to the post-test that consisted of a vocabulary test on the assigned topic (same 10 questions as in the pre-test in shuffled order) and an essay writing assignment (100+ words).
7. **Delay-test** One week after the post-test, participants were invited to take a delay-test which consisted of the vocabulary test as the post-test in different question order.

4.4 Participants

We conducted our user study on the *Prolific Academic*[4] platform. The required number of participants was determined by a statistical power analysis conducted with a significance level of $\alpha = 0.05$, a power of $1 - \beta = 0.80$, an expected effect size of 0.25 and a group size of 4 using the software *GPower* [30]. This gave a minimum required number of $n = 136$ participants. To ensure the response quality, we only recruited native English-speaking participants within the age range of 18 to 51 with a minimum of 95% approval rate, at least 100 successful task submissions, and at least a high-school level of education. The entire study lasted for around 35 minutes. We paid each participant GBP £5 for the study. Overall, 178 participants completed the post-test; we rejected 18 of them because of a lack of attention (over 5 minutes of no activity in the browser tab) in the search phase, which led to 160 valid participants. We further paid £1 bonus for participants who took the delay-test after one week, and 144 valid participants returned and completed the delay-test. Among the 144 participants (77 male, 67 female), the median age is 34.5 (min. 20, max. 51). Forty reported a high school degree as the highest education degree, 17 reported a community college degree, 58 reported an undergraduate degree, 25 reported a graduated degree, and 4 reported doctorate degrees. Table 3 reports the number of participants over each topic and each test condition. The 144 participants were evenly distributed among the topics, each with a participant count ranging from 19 to 23. Table 3 also shows the average number of queries over each topic, which ranges from 3.26 to 4.65, indicating that our participants actively engaged in the search phase.

[4]https://app.prolific.co

4.5 Metrics

4.5.1 Learning Gains. In the pre-, post-, and delay–tests, we asked our participants to self-assess their knowledge levels on a set of vocabulary terms. In line with [11, 50, 59, 60], we evaluated the study participants' knowledge of a term with the *Vocabulary Knowledge Scale (VKS)* [73] across four levels:

1 *I don't remember having seen this term/phrase before.*
2 *I have seen this term/phrase before, but I don't think I know what it means.*
3 *I have seen this term/phrase before, and I think it means …*
4 *I know this term/phrase. It means …*

For both levels *(3)* and *(4)*, we further asked participants to write down the meaning of the vocabulary term in their own words that we can use to judge the quality and reliability of the self-assessment. To reduce the question priming effects, participants did not know that vocabulary terms asked in the pre-test would be asked again in the post-test. Following earlier works [11, 50, 60], we first rescored the knowledge level self-assessments to $0-2$. We assigned a score of 0 to knowledge level *(1)* and *(2)*, a score of 1 to knowledge level *(3)*, and a score of 2 to knowledge level *(4)*. Then we evaluated the learning gain with *Realized Potential Learning (RPL)* [18, 50, 63], which is the absolute knowledge gains (ALG) measured by the number of new vocabulary terms learned (a score change of 0 to 1 or 2 from pre-test to subsequent tests) and the number of vocabulary terms they became more confident at (a score change of 1 to 2) normalized by the maximum possible learning gain:

$$ALG = \frac{1}{n}\sum_{i=1}^{n} \max(0, vks^{x}(v_i) - vks^{pre}(v_i)) \tag{1}$$

$$RPL = \frac{ALG}{\frac{1}{n}\sum_{i=1}^{n} 2 - vks^{pre}(v_i)} \tag{2}$$

where $vks^{pre}(v_i)$ is the rescored knowledge level of vocabulary v_i in pre-test; $vks^{x}(v_i), x \in \{\text{post}, \text{delay}\}$ is the rescored knowledge

Table 3: Overview of topics and corresponding vocabulary terms chosen for learning tasks, as well as number of participants and other associated statistics (± represents the standard deviation) over topics. Two-way ANOVA tests revealed no significant differences in the average number of queries ($F(6, 132) = 0.839, p = 0.542$).

	Ethics	**Genetically modified organism**	**Noise-induced hearing loss**	**Radiocarbon dating considerations**	**Business cycle**	**Irritable bowel syndrome**	**Theory of mind**
Vocabulary Terms	anarchist ethics, descriptive ethics, normative ethics, relational ethics, virtue ethics, ethical resistance, consequentialism, epicurean ethics, ethics feasible, ethics spheres	transgenic, genomes, selective breeding, microinjection enzyme, chromosome, plasmid, myxoma, kanamycin, severe combined immunodeficiency, Leber's congenital amaurosis	acoustic trauma, discomfort threshold, cochlear damage, audiogram, overstimulation of hair cells, noise conditioning, excitotoxicity, OSHA, sensorineural hearing loss, tinnitus, Threshold shift	carbon exchange reservoir, isotopic fractionation, polarity excursion, carbonate, geomagnetic reversals, mass spectrometry, upwelling, radiocarbon, neutrons, photosynthesis pathways	economic cycles, distribution cycles, swing cycle, wage cycle, marxist model, endogenous causes, friedman, capital profitability, model recession, austrian school	bifidobacteria infantis, mesalazin, bile acid malabsorption, selective serotonin reuptake inhibitors, Gut-brain axis, antidepressants, laxatives, probiotics, celiac disease, epithelial barrier	asperger syndrome, theory of mindreading, attentional reorienting, mind development, mind autism, hyperactivity, perspective experiment, intentionality, perception, belief
Number of participants	21	20	21	19	19	21	23
$\mathbf{Q}_{\mathbf{none}}$	3	4	6	5	5	4	5
$\mathbf{Q}_{\mathbf{synthesis}}$	7	6	5	3	5	4	7
$\mathbf{Q}_{\mathbf{random}}$	6	6	4	5	4	5	6
$\mathbf{Q}_{\mathbf{term}}$	5	4	6	6	5	8	5
Average number of queries	4.43(±4.20)	4.65(±3.48)	3.48(±2.34)	4.00(±3.64)	4.42(±3.44)	3.71(±2.63)	3.26(±3.52)
Median number of queries	3.00	3.50	3.00	2.00	3.00	3.00	2.00

level of vocabulary v_i in post- or delay-test. $vks(v_i) \in \{0, 1, 2\}$ and n is the number of vocabulary items under the tested topic.

4.5.2 Self-assessment Quality. In order to determine the quality of vocabulary knowledge self-assessments, we sampled $approx10\%$ of term definitions of knowledge levels (3) and (4) from both the pre- and post-tests (specifically, 40 from the pre-test and 60 from the post-test) written by participants. We tasked two experts to label these definitions as either *correct*, *partially correct* or *incorrect* keeping in mind that the definitions were written by topical novices. Based on the expert labels, among definitions of knowledge level (3), 20% were correct, 68% were partially correct, and the remaining 12% were incorrect. Among the definitions of knowledge level (4), 70% were correct, 24% were partially correct, and the remaining 6% were incorrect. Based on the low incorrect rate, we consider the self-assessment reliable.

4.5.3 Automatic Assessment. Another way to scale up the assessment of our participants' definitions is to rely on large-scale language models (LLMs). State-of-the-art LLMs such as GPT-3.5 or GPT-4 [51] have reportedly achieved human-level performance on various complex natural language tasks. To evaluate the influence of uncertainty in self-assessment, we evaluated all definitions by prompting GPT-3.5. Based on GPT-3.5's output, we categorized *partially correct* term definitions as knowledge level (3) and *correct* term definitions as knowledge level (4). Incorrect term definitions were designated as level (2). We conducted our data analyses with both the self-assessment knowledge levels and the knowledge levels as determined by GPT-3.5. The trends and statistical outcomes do not differ between self-assessment and GPT-3.5-based assessment. As an example, for the learning gain metric RPL, the scores for $\mathbf{Q}_{\text{none}}$, $\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$ conditions are 0.22(±0.16), 0.14(±0.13), 0.13(±0.14), 0.20(±0.14) respectively. **Thus, in the remainder of this paper, we report the learning gain evaluation based on the self-assessed vocabulary knowledge levels only.**

4.5.4 Essay Quality. In addition to RPL, we evaluated knowledge expressed in participants' essays with two additional measures as learning indicators: F-Fact and T-Depth, following [60, 75]. Concretely, F-Fact represents the number of individual facts in an essay, and T-Depth represents the extent to which each subtopic is covered. We manually annotated the written essays for both measures. For F-Fact, the annotators were required to identify topic-related facts, e.g., *"GMOs have not shown to be any more harmful"*, and count the number of facts in the entire essay. For T-Depth, annotators scored the essay concerning each subtopic on a scale of 0 to 3, where 0 represented not covered, and 3 indicated the essay covered the subtopic with great focus. The overall T-Depth score is the average of all subtopic T-Depth scores in the same topic. Five annotators (authors of this paper) divided 160 essays among themselves. Twenty essays were annotated by all annotators, achieving a Pearson correlation of 0.73 for *F-Fact* and 0.75 for *T-Depth*, indicating high inter-annotator agreement for the metrics.

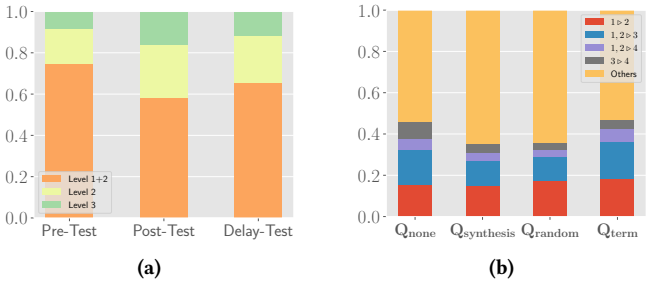
4.5.5 Behaviour Metrics. The engagement in the search process is usually correlated with the learning outcome. Previous works [17, 59, 68, 76] have investigated a series of effective behaviour metrics as proxy measures for learning. Following prior research, we extracted seven types of search and reading behaviour from our collected search logs: (i) the **number of queries** a participant formulates; (ii) the **number of unique documents** a participant viewed; (iii) the **number of snippets** a participant viewed; (iv) the **average time of between queries**; (v) the **average time between documents**; (vi) the **average document dwell time**; (vii) the **number of mouse scrolls** over the opened documents.

4.5.6 Answer Quality. To examine whether participants indeed engaged with the adjunct questions, we evaluated participant-written answer quality of the factoid questions with EM (Exact Match) score, which measures the percentage of answers that match exactly with ground-truth answers, and (macro-averaged) F1 score [31, 55], which treats all answers as bags of tokens and calculate the average overlap between the participants' answers and the ground truth answer. We found the F1 scores of answers to $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ were 0.589 and 0.408, and the EM scores of answers to $\mathbf{Q}_{\text{random}}$ and $\mathbf{Q}_{\text{term}}$ were 0.523 and 0.322, respectively. These results first confirmed that participants indeed engaged with the questions. However, participants cannot always find the exact answer spans (The EM score was lower than the F1 score in both conditions). Both F1 and EM scores of $\mathbf{Q}_{\text{term}}$ answers were lower than $\mathbf{Q}_{\text{random}}$

answers, indicating that questions on vocabulary terms were more challenging to answer compared to questions on random facts from the document.

5 RESULTS

In this section, we discuss the results of our user study. As a sanity check, we first analyze participants' overall learning gains. Figure 3a reports the distribution of knowledge levels reported in pre-, post-, and delay-tests. Participants marked fewer vocabulary terms as knowledge levels 1 or 2 and more as knowledge levels 3 and 4 in post- and delay-tests than the pre-test, which shows that participants learned both short-term (post-test) and long-term (delay-test) vocabulary knowledge over the assigned topics in the task. Furthermore, Figure 3b shows detailed knowledge state transitions on each condition from pre-test to post-test. Although the assessment on most vocabulary terms (> 50%) remained unchanged and transitions among lower knowledge levels accounted for most learning gains, participants did achieve learning gains in all conditions. These results, together with the evaluation of participants' self-assessment quality (Section 4.5.2) and the quality of answers to the adjunct questions (Section 4.5.6), validate our system and experimental design. On average, participants were indeed actively engaged and learning throughout the study.



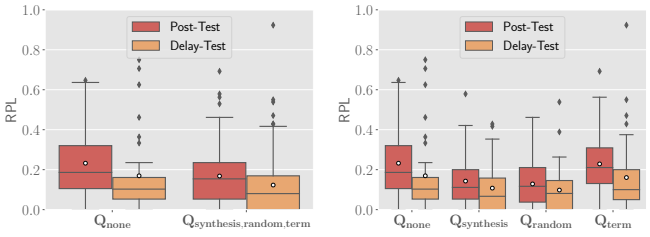
(a) (b)

Figure 3: (a) Distributions of vocabulary knowledge levels in the pre-test, post-test, and delay-test. (b) The fraction of vocabulary knowledge changes from pre-test to post-test.

We now present study results in line with the research questions. Table 4 presents the main results. We conducted two-way ANOVA tests on these measures, considering the assigned topics and the conditions as factors, and examined the main effects with $\alpha = 0.05$. TukeyHSD pairwise tests were used for post-hoc analysis.

5.1 RQ1: Adjunct Question Effects in SAL

5.1.1 Effects of Adjunct Questions on Participants' Search Behaviour. In our user study, participants were required to learn one topic by searching and reading for at least 20 minutes. The average document dwell time (Row **XIV** in Table 4) of participants who received adjunct questions was significantly longer than $\mathbf{Q}_{\text{none}}$ ($p < 0.05$). As a consequence of the longer dwell time, we also observed the number of queries (Row **IX**, $F(3, 116) = 6.70, p = 3\times10^{-4}$), the number of unique documents (Row **X**, $p = 2\times10^{-6}$), and the number of unique snippets (Row **XI**, $p = 0.001$) that participants viewed to be significantly lower than participants in $\mathbf{Q}_{\text{none}}$. The average time between queries of participants with adjunct questions (ranging from 636 s to 741 s) was significantly longer than $\mathbf{Q}_{\text{none}}$ participants (Row **XII**, $p = 3.7 \times 10^{-4}$). In addition, we also measured participants' in-document mouse activities, i.e., the number of scrolls while reading one document (Row **XV**). We observed that participants had more scrolls when presented with adjunct questions, indicating more concentrated reading behaviour. *These results confirm that adjunct questions significantly impact participants' behaviour, which is consistent with findings from [68].*



(a) (b)

Figure 4: Distribution of post-test and delay-test RPL scores (a) w/ vs. w/o adjunct questions, (b) with all conditions.

5.1.2 Effects of Adjunct Questions on Participants' Learning Gains. Recall that we evaluated participants' learning outcomes with RPL. Figure 4a shows that in both the post- and delay-test, the RPL of $\mathbf{Q}_{\text{none}}$ participants was higher than that of participants who received adjunct questions ($\mathbf{Q}_{\text{synthesis}}$, $\mathbf{Q}_{\text{random}}$, and $\mathbf{Q}_{\text{term}}$) (Post-test: $M. = 0.23$ vs. $M. = 0.17, p = 0.026$, Delay-test: $M. = 0.17$ vs. $M. = 0.12, p = 0.14$, where $M.$ represents the Mean value). Figure 4b shows a detailed comparison broken down to all conditions. The $\mathbf{Q}_{\text{none}}$ and the $\mathbf{Q}_{\text{term}}$ participants had similar short-term retention ($M. = 0.23$ vs. $M. = 0.23$). Both were significantly higher than the $\mathbf{Q}_{\text{random}}$ condition ($M. = 0.23$ vs $M. = 0.13, p = 0.02$) and higher than $\mathbf{Q}_{\text{synthesis}}$ ($M. = 0.24$ vs $M. = 0.14, p = 0.06$). The delay-test RPL (Row **IV**) reflects the long-term learning outcomes. $\mathbf{Q}_{\text{random}}$ exhibited the worst results; $\mathbf{Q}_{\text{term}}$ was close to $\mathbf{Q}_{\text{none}}$. Previous work like [68] showed that participants spent substantially more time reading the same reading materials when presented with adjunct questions. Recall that participants had limited task time, and in the adjunct question conditions, participants read significantly fewer documents, which can partly explain the negative effects of adjunct questions. This is also aligned with an earlier classroom study of adjunct questions [35], where the length of the task time is an essential factor in learning outcomes.

To sum up, our study revealed that participants who received adjunct questions exhibited more fine-grained reading behaviour but had lower retention. However, when appropriate questions were posed, these participants achieved comparable short-term and long-term learning gains while reading significantly fewer documents. This highlights the importance of understanding learners' knowledge requirements and time constraints for presenting adjunct questions.

Table 4: Mean (± standard deviations) of evaluation metrics across all participants in each condition. † denotes the two-way ANOVA significance, while $\mathcal{N}$,$\mathcal{S}$,$\mathcal{R}$,$\mathcal{T}$ indicate post-hoc significance (TukeyHSD pairwise test, p<0.05) over the four conditions Q_{none}, $Q_{synthesis}$, Q_{random}, and Q_{term}, respectively.

	Measure	Q_{none}	$Q_{synthesis}$	Q_{random}	Q_{term}
I.	Number of participants	32	37	36	39
II.	Search phase	19m47s(±1m46s)	21m5s(±3m16s)	20m6s(±2m30s)	20m18s(±2m56s)
III.	Post-Test RPL†	0.23(±0.18)$^{\mathcal{R}}$	0.14(±0.14)	0.13(±0.12)$^{\mathcal{NT}}$	0.23(±0.15)$^{\mathcal{R}}$
IV.	Delay-Test RPL	0.17(±0.20)	0.11(±0.12)	0.10(±0.12)	0.16(±0.19)
V.	Flesch score	52.00(±12.33)	49.67(±17.06)	52.56(±13.23)	56.49(±13.42)
VI.	T-Depth	0.87(±0.35)	0.91(±0.48)	0.78(±0.47)	0.81(±0.41)
VII.	F-Fact	13.88(±7.82)	12.68(±6.38)	10.16(±7.28)	12.68(±8.57)
VIII.	Fraction of topical terms used by essays	0.05(±0.04)	0.04(±0.03)	0.03(±0.03)	0.04(±0.02)
IX.	Number of queries†	6.09(±3.90)$^{\mathcal{SRT}}$	3.49(±2.80)$^{\mathcal{N}}$	3.11(±2.78)$^{\mathcal{N}}$	3.49(±3.16)$^{\mathcal{N}}$
X.	Number of unique documents viewed†	13.44(±6.65)$^{\mathcal{SRT}}$	7.49(±3.01)$^{\mathcal{N}}$	9.47(±5.14)$^{\mathcal{N}}$	9.10(±3.89)$^{\mathcal{N}}$
XI.	Number of snippets†	45.09(±21.16)$^{\mathcal{SRT}}$	31.57(±17.55)$^{\mathcal{N}}$	32.47(±16.61)$^{\mathcal{N}}$	28.38(±15.70)$^{\mathcal{N}}$
XII.	Average time between queries (secs.)†	379.97(±329.66)$^{\mathcal{SRT}}$	685.47(±368.11)$^{\mathcal{N}}$	740.67(±375.35)$^{\mathcal{N}}$	615.45(±380.46)$^{\mathcal{N}}$
XIII.	Average time between documents (secs.)	18.78(±16.89)	20.20(±13.70)	20.40(±19.73)	19.85(±20.87)
XIV.	Average document dwell time(s)†	73.99(±35.46)$^{\mathcal{ST}}$	182.89(±149.15)$^{\mathcal{NRT}}$	128.21(±54.73)$^{\mathcal{S}}$	128.93(±59.32)$^{\mathcal{NS}}$
XV	Number of scrolls†	14.62(±18.76)$^{\mathcal{S}}$	98.18(±94.79)$^{\mathcal{NRT}}$	38.99(±43.97)$^{\mathcal{S}}$	36.30(±28.86)$^{\mathcal{S}}$
XVI.	Average number of non-stopwords in answers†	–	6.75(±4.50)$^{\mathcal{RT}}$	0.82(±1.02)$^{\mathcal{S}}$	1.15(±1.38)$^{\mathcal{S}}$
XVII.	Average reading time before answering (secs.)	–	116.43(±120.90)	98.32(±46.91)	89.73(±48.79)
XVIII.	Average time to create answers (secs.)†	–	36.55(±24.17)$^{\mathcal{RT}}$	9.82(±8.79)$^{\mathcal{S}}$	13.49(±16.28)$^{\mathcal{S}}$
XIX.	F1 score†	–	–	0.58(±0.20)$^{\mathcal{T}}$	0.39(±0.23)$^{\mathcal{R}}$
XX.	EM score†	–	–	0.51(±0.24)$^{\mathcal{T}}$	0.30(±0.26)$^{\mathcal{R}}$

5.2 RQ2: Factors that Influence Automatically Generated Adjunct Questions' Effects

5.2.1 Impacts of Question Types. Syed et al. [68] found that participants spent more time reading with additional synthesis questions while having similar learning gains with those who received only factoid questions. As shown in Row **XIV** of Table 4, compared to participants in factoid question conditions (Q_{random} and Q_{term}), $Q_{synthesis}$ participants had significantly longer average document dwell time (183s, $p < 0.05$). Similarly, $Q_{synthesis}$ participants executed significantly more scrolls (Row **XV**, $M. = 98$, $p < 10^{-4}$) and spent more time reading before answering the adjunct question (Row **XVII**, $M. = 116s$ vs. $M. = 98s$ and 90s respectively) than participants who received factoid questions. Additionally, participants in the $Q_{synthesis}$ condition produced significantly longer answers for adjunct questions (Row **XVI**, $p < 10^{-18}$) and spent the longest time writing their answers (Row **XVIII**, $p < 10^{-9}$). These results indicate that compared to factoid questions, the generated synthesis questions cause more cognitive burden. $Q_{synthesis}$ participants have to spend more time reading, rewinding, and writing answers, which aligns with the previous work [68].

Regarding the learning outcomes, $Q_{synthesis}$ participants had similar RPL to Q_{random} in both post-test ($M. = 0.14$ vs. $M. = 0.13$) and delay-test ($M. = 0.11$ vs. $M. = 0.10$). In addition, we also measured participants' learning with essay writing (at least 100 words) in the post-test. Specifically, we consider T-Depth (measuring the number of subtopics covered) and F-Fact (measuring the number of atomic facts). As seen in Table 4 (Row **VI** and Row **VII**), $Q_{synthesis}$ participants exhibited the highest T-Depth score among all adjunct question conditions. These results showed that although essays created by participants in $Q_{synthesis}$ were the most difficult to read (with the lowest Flesch reading ease score of 49.67), they provided better topic coverage and a comparable number of facts. Thus, we conclude that compared with factoid questions, synthesis questions may cause a higher cognitive burden and higher performance on tests (i.e., essay writing) that require higher cognitive complexity than factoid questions regarding random text spans.

5.2.2 Impacts of Question Target Selection. As target selection is an essential procedure for generating questions, especially factoid questions, we investigate the effects of question target selection via the conditions Q_{random} and Q_{term}. Recall that the answers (for which to generate questions) in Q_{random} were text spans extracted from the document and the answers for Q_{term} were the vocabulary terms of the assigned topic. To this end, we collected 377 document viewings in condition Q_{term}, 63.4% of which targeted the assigned topic's terms. 58.8% of all topic terms were covered. As seen in rows from **IX** to **XV** of Table 4, participants in Q_{random} and Q_{term} did not show significant differences in the activity measures, although Q_{random} spent more time between queries on average. When it came to answering the questions, we found Q_{term} participants spent less time reading before writing answers (Row **XVII**, $M. = 89.73$ vs. $M. = 98.32$, $p = 0.89$) and spent longer time writing answers (Row **XVIII**, $M. = 13.49$ vs. $M. = 9.82$, $p = 0.64$) than Q_{random}, but these differences were also not statistically significant. In contrast, Q_{term} participants' answers to adjunct questions showed significantly lower quality in terms of the F1 score (Row **XIX**, $M. = 0.39$ vs. $M. = 0.58$, $p < 10^{-4}$) and EM score (Row **XX**, $M. = 0.3$ vs. $M. = 0.58$, $p < 10^{-4}$). This may be due to the different complexity of the target answers. Vocabulary terms of the assigned topic tend to be more complex than the randomly chosen answer spans, such as named entities in the document. Notably, Q_{term} participants had better short-term learning outcomes (Row **III**, Post-test RPL, $M. = 0.23$ vs. $M. = 0.13$, $p = 0.02$) and long-term retention (Row **IV**, Delay-test RPL, $M = 0.16$ vs. $M = 0.10$, $p = 0.06$) than Q_{random} participants. We also found that Q_{term} participants had better yet not significant essay quality in all evaluated measures compared to Q_{random} participants. These observations suggest that compared

with random target answer selection, guiding the users according to their learning goals achieves significantly better learning gains despite similar observed search behaviour, indicating the importance of learning goal-aware adaptive AQG for adjunct questions.

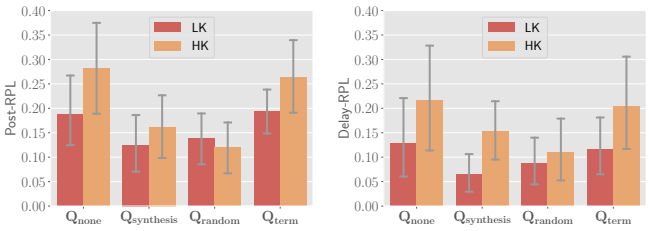


Figure 5: Distribution of post-test (a) and delay-test (b) RPL scores of HK and LK participants with all conditions.

5.2.3 Impacts of Prior-Knowledge. Participants' prior knowledge may influence their behaviour, like the reading time [14] and their ability to identify the answer without reading. Thus, the effects of adjunct questions are sensitive to participants' prior knowledge levels [68] and may cause contrasting effects. In this paper, we considered a participant as high-knowledge (HK) for a topic if her average pre-test score was higher than the median and otherwise low-knowledge (LK). We classified 71 participants as HK participants and 73 as LK participants. Figure 5 compares the RPL of HK and LK participants in each condition in post-test (Figure 5a) and delay-test (Figure 5b). On average, HK participants exhibited higher RPL scores in all conditions during both the post-test and delay-test except $\mathbf{Q}_{\text{random}}$ in the post-test. Specifically, in the delay-test, $\mathbf{Q}_{\text{synthesis}}$ HK participants had significantly higher RPL than the LK group ($M. = 0.15$ vs. $M. = 0.06$, $p = 0.044$). Moreover, compared with the RPL decrease from the post-test to the delay-test in other conditions, $\mathbf{Q}_{\text{synthesis}}$ HK participants show a slighter RPL decrease ($0.16 \Rightarrow 0.15$). These results indicate that synthetic questions that require higher-level cognition lead to better long-term retention for more knowledgeable learners than other conditions.

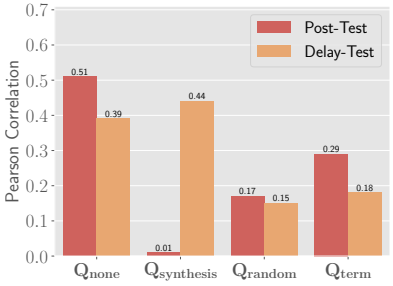


Figure 6: The Pearson correlation values between RPL and the average pre-test score in post-test and delay-test.

We further investigated the correlation between the learning outcomes (both short-term and long-term) measured by RPL and participants' prior-knowledge measured by the pre-test vocabulary knowledge evaluation. Figure 6 shows the Pearson correlation scores. First, across all conditions, both post- and delay-test RPL scores were positively related to participants' prior knowledge. Furthermore, we found that $\mathbf{Q}_{\text{none}}$ participants' post-test RPL was strongly correlated with their prior-knowledge. However, adjunct questions mitigated the correlation, particularly in $\mathbf{Q}_{\text{synthesis}}$ condition, where there was no correlation between the post-test RPL and prior-knowledge. Lastly, we noted that the correlation with prior-knowledge was generally weaker in the delay-test than in the post-test, except in the $\mathbf{Q}_{\text{synthesis}}$ condition, where in contrast participants exhibited a much stronger correlation in the delay-test than in the post-test. These findings indicate the importance of adapting different AQG strategies based on learners' prior-knowledge levels.

6 LIMITATIONS AND FUTURE WORK

We note some limitations stemming from the study design and the result assessments of our study that indicate potential directions for future research. First, we generate questions with the PAQ framework. As Large Language Models (LLMs) have demonstrated significantly better zero-shot and few-shot text generation quality, it is natural to extend this research to study the effects of applying LLMs for adjunct question generation in the SAL scenario. Furthermore, we adopted the vocabulary learning task and the essay writing task for evaluating learning outcomes. We note the deviation between conclusions drawn from vocabulary learning and essay writing tasks. The vocabulary learning task may limit the ability to assess the understanding of deeper learning levels, which encourages further research on knowledge assessment methods. Last, the length of the task time is an essential factor in learning outcomes, but in this study, all participants in the experiment spent around 20 minutes in the search phase. As we can observe in Table 4, participants who received adjunct questions issued fewer queries and viewed fewer documents than those in the control condition. This may explain the findings that participants in the control condition scored higher than those who received questions.

7 CONCLUSIONS

This paper explored the effects of automatically generated adjunct questions in the complex search as learning scenario through a user study. The empirical results confirm previous findings—adjunct questions significantly influence participants' behaviour and learning outcomes, though in our study these effects vary across different conditions. We found evidence that with adjunct questions, participants were more engaged with the search results than those without adjunct questions. As a potential consequence of longer reading time, adjunct questions may negatively affect learning gains if the learning time is the same across all conditions. Furthermore, our results demonstrate the importance of adopting different types of adjunct questions for learning tasks with different cognitive complexity. Selecting targeting answers for adjunct questions according to participants' learning goals can significantly improve participants' learning outcomes. Lastly, we found participants' prior-knowledge had essential impacts on their learning gains, especially when they were posed with adjunct questions that required higher cognitive levels. Adjunct questions may mitigate the correlation between learning outcomes and the prior-knowledge.

REFERENCES

[1] Alireza Ahadi, Abhay Singh, Matt Bower, and Michael Garrett. 2022. Text mining in education—A bibliometrics-based systematic review. *Education Sciences* 12, 3 (2022), 210.

[2] Richard C Anderson and W Barry Biddle. 1975. On asking people questions about what they are reading. In *Psychology of learning and motivation*. Vol. 9. Elsevier, 89–132.

[3] Kumaripaba Athukorala, Alan Medlar, Antti Oulasvirta, Giulio Jacucci, and Dorota Glowacka. [n. d.]. Beyond Relevance: Adapting Exploration/Exploitation in Information Retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (2016-03-07) *(IUI '16)*. Association for Computing Machinery, 359–369.

[4] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Songhao Piao, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In *Preprint*.

[5] Marcia J. Bates. 1989. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review* 13, 5 (Jan. 1989), 407–424.

[6] Nicholas J. Belkin. 1990. The Cognitive Viewpoint in Information Science. *Journal of Information Science* 16, 1 (Feb. 1990), 11–15.

[7] Benjamin Samuel Bloom, Peter Airasian, Kathleen Cruikshank, Richard Mayer, Paul Pintrich, James Raths, and Merlin Wittrock. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Pearson.

[8] Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. 819–826.

[9] Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable Educational Question Generation with Pre-trained Language Models. *arXiv preprint arXiv:2305.07871* (2023).

[10] Katriina Byström and Kalervo Järvelin. 1995. Task Complexity Affects Information Seeking and Use. *Information Processing & Management* 31, 2 (March 1995), 191–213.

[11] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2021. Searching to learn with instructional scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 209–218.

[12] Shana K Carpenter, Harold Pashler, John T Wixted, and Edward Vul. 2008. The effects of tests on learning and forgetting. *Memory & Cognition* 36, 2 (2008), 438–448.

[13] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. LearningQ: a large-scale dataset for educational question generation. In *Twelfth International AAAI Conference on Web and Social Media*.

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

[15] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. [n. d.]. Personalizing Web Search Results by Reading Level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (2011-10-24) *(CIKM '11)*. Association for Computing Machinery, 403–412.

[16] Kevyn Collins-Thompson, Preben Hansen, and Claudia Hauff. 2017. Search as learning (dagstuhl seminar 17092). In *Dagstuhl reports*, Vol. 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[17] Kevyn Collins-Thompson, Soo Young Rieh, Carl C Haynes, and Rohail Syed. 2016. Assessing learning outcomes in web search: A comparison of tasks and query strategies. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval*. 163–172.

[18] Henri G Colt, Mohsen Davoudi, Septimiu Murgu, and Nazanin Zamanian Rohani. 2011. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy* 25, 1 (2011), 207–216.

[19] Bidyut Das, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2021. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning* 16, 1 (2021), 1–15.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Laura Dietz. 2019. TREC CAR Y3: Complex Answer Retrieval Overview.

[22] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *Comput. Surveys* 55, 8 (2022), 1–38.

[23] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*. 13042–13054.

[24] Michele M Dornisch. 2012. Adjunct questions: Effects on learning. *Encyclopedia of the Sciences of Learning* (2012), 128–129.

[25] Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106* (2017).

[26] John Dunlosky, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham. 2013. Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 14, 1 (Jan. 2013), 4–58.

[27] Alexander R Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. *arXiv preprint arXiv:2004.11892* (2020).

[28] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. *arXiv preprint arXiv:1907.09190* (2019).

[29] Warren Fass and Gary M. Schumacher. 1978. Effects of Motivation, Subject Activity, and Readability on the Retention of Prose Materials. *Journal of Educational Psychology* 70 (1978), 803–807.

[30] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[31] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. *arXiv preprint arXiv:1910.09753* (2019).

[32] Craig R Fox and Jonathan Levav. 2000. Familiarity bias and belief reversal in relative likelihood judgment. *Organizational Behavior and Human Decision Processes* 82, 2 (2000), 268–292.

[33] Xiaoran Fu, K Lokesh Krishna, and R Sabitha. 2021. Artificial Intelligence Applications with e-Learning System for China's Higher Education Platform. *Journal of Interconnection Networks* (2021), 2143016.

[34] 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.

[35] Christiaan Hamaker. 1986. The effects of adjunct questions on prose learning. *Review of educational research* 56, 2 (1986), 212–242.

[36] Michael Heilman. 2011. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University* 195 (2011).

[37] Cheryl I Johnson and Richard E Mayer. 2009. A testing effect with multimedia learning. *Journal of Educational Psychology* 101, 3 (2009), 621.

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

[39] Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education* 30 (2020), 121–204.

[40] Che-Hao Lee, Tzu-Yu Chen, Liang-Pu Chen, Ping-Che Yang, and Richard Tzong-Han Tsai. 2018. Automatic question generation from children's stories for companion chatbot. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 491–494.

[41] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[42] Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics* 9 (2021), 1098–1115.

[43] Chang Liu, Jacek Gwizdka, Jingjing Liu, Tao Xu, and Nicholas J. Belkin. [n. d.]. Analysis and Evaluation of Query Reformulations in Different Task Types. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47* (2010-10-22) *(ASIS&T'10)*. American Society for Information Science, 1–10.

[44] Hanrui Liu, Chang Liu, and Nicholas J Belkin. 2019. Investigation of users' knowledge change process in learning-related search tasks. *Proceedings of the Association for Information Science and Technology* 56, 1 (2019), 166–175.

[45] Hanrui Liu, Chang Liu, and Nicholas J. Belkin. 2019. Investigation of Users' Knowledge Change Process in Learning-Related Search Tasks. *Proceedings of the Association for Information Science and Technology* 56, 1 (2019), 166–175.

[46] Owen HT Lu, Anna YQ Huang, Danny CL Tsai, and Stephen JH Yang. 2021. Expert-Authored and Machine-Generated Short-Answer Questions for Assessing Students Learning Performance. *Educational Technology & Society* 24, 3 (2021), 159–173.

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

[48] David Maxwell and Claudia Hauff. 2021. LogUI: contemporary logging infrastructure for web-based experiments. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 525–530.

[49] Ruslan Mitkov et al. 2003. Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications*

using natural language processing. 17–22.
[50] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. 2018. Contrasting search as a learning activity with instructor-designed learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 167–176.
[51] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023).
[52] Sindunuraga Rikarno Putra, Kilian Grashoff, Felipe Moraes, and Claudia Hauff. 2018. On the Development of a Collaborative Search System.. In *DESIRES*. 76–82.
[53] Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation. *arXiv preprint arXiv:2104.08006* (2021).
[54] Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. [n. d.]. Conversational Interfaces for Search As Learning. In *Proceedings of the First International Workshop on Investigating Learning During Web Search* (2020-10-19/2020-10-20) *(IWILDS 2020)*. 4.
[55] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).
[56] Ernst Z Rothkopf. 1965. Some theoretical and experimental approaches to problems in written instruction. *Learning and the educational process. Chicago: Rand McNally* 965 (1965).
[57] Ernst Z Rothkopf. 1966. Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal* 3, 4 (1966), 241–249.
[58] Jean-François Rouet and Eduardo Vidal-Abarca. 2002. Mining for meaning: Cognitive effects of inserted questions in learning from scientific text. *The psychology of science text comprehension* (2002), 417–436.
[59] Nirmal Roy, Felipe Moraes, and Claudia Hauff. 2020. Exploring users' learning gains within search sessions. In *Proceedings of the 2020 conference on human information interaction and retrieval*. 432–436.
[60] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the highlight: incorporating active reading tools in a search as learning environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 229–238.
[61] Sara Salimzadeh, David Maxwell, and Claudia Hauff. [n. d.]. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (2021-08-31) *(ICTIR '21)*. Association for Computing Machinery, 63–72.
[62] Amy M. Shapiro. 1998. Promoting Active Learning: The Role of System Structure in Learning From Hypertext. *Human–Computer Interaction* 13, 1 (March 1998), 1–35.
[63] John L Shefelbine. 1990. Student factors related to variability in learning word meanings from context. *Journal of Reading Behavior* 22, 1 (1990), 71–97.
[64] Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. *arXiv preprint arXiv:2106.04262* (2021).
[65] Tim Steuer, Anna Filighera, and Christoph Rensing. 2020. Remember the facts? Investigating answer-aware neural question generation for text comprehension. In *International Conference on Artificial Intelligence in Education*. Springer, 512–523.
[66] Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. 2022. Educational Automatic Question Generation Improves Reading Comprehension in Non-native Speakers: A Learner-Centric Case Study. *Frontiers in Artificial Intelligence* 5 (2022).
[67] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 555–564.
[68] Rohail Syed, Kevyn Collins-Thompson, Paul N Bennett, Mengqiu Teng, Shane Williams, Dr Wendy W Tay, and Shamsi Iqbal. 2020. Improving learning outcomes with gaze tracking and automatic question generation. In *Proceedings of The Web Conference 2020*. 1693–1703.
[69] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. [n. d.]. ScentBar: A Query Suggestion Interface Visualizing the Amount of Missed Relevant Information for Intrinsically Diverse Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2016-07-07) *(SIGIR '16)*. Association for Computing Machinery, 405–414.
[70] Rachel Van Campenhout, Nick Brown, Bill Jerome, Jeffrey S Dittel, and Benny G Johnson. 2021. Toward effective courseware at scale: Investigating automatically generated questions as formative practice. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*. 295–298.
[71] Johannes Von Hoyer, Anett Hoppe, Yvonne Kammerer, Christian Otto, Georg Pardi, Markus Rokicki, Ran Yu, Stefan Dietze, Ralph Ewerth, and Peter Holtz. 2022. The Search as Learning Spaceship: Toward a Comprehensive Model of Psychological and Technological Facets of Search as Learning. *Frontiers in Psychology* 13 (2022), 827748–827748.
[72] Zichao Wang, Andrew S Lan, Weili Nie, Andrew E Waters, Phillip J Grimaldi, and Richard G Baraniuk. 2018. QG-net: a data-driven question generation model for educational content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. 1–10.
[73] Marjorie Wesche and T Sima Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53, 1 (1996), 13–40.
[74] Ryen W. White, Susan T. Dumais, and Jaime Teevan. [n. d.]. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (2009-02-09) *(WSDM '09)*. Association for Computing Machinery, 132–141.
[75] Mathew J Wilson and Max L Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 291–306.
[76] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting user knowledge gain in informational search sessions. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 75–84.