UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry

Matthias Stadler, Maria Bannert, Michael Sailer

Angaben zur Veröffentlichung / Publication details:

Stadler, Matthias, Maria Bannert, and Michael Sailer. 2024. "Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry." *Computers in Human Behavior* 160: 108386. https://doi.org/10.1016/j.chb.2024.108386.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Computers in Human Behavior 160 (2024) 108386



Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh



Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry

Matthias Stadler [a,*], Maria Bannert [b], Michael Sailer [c]

[a] Institute of Medical Education, LMU University Hospital, LMU Munich, Germany
[b] Chair for Teaching and Learning with Digital Media, Technical University of Munich, Germany
[c] Learning Analytics and Educational Data Mining, University of Augsburg, Augsburg, Germany

ARTICLE INFO

Handling editor: Nicolae Nistor

ABSTRACT

This study explores the cognitive load and learning outcomes associated with using large language models (LLMs) versus traditional search engines for information gathering during learning. A total of 91 university students were randomly assigned to either use ChatGPT3.5 or Google to research the socio-scientific issue of nanoparticles in sunscreen to derive valid recommendations and justifications. The study aimed to investigate potential differences in cognitive load, as well as the quality and homogeneity of the students' recommendations and justifications. Results indicated that students using LLMs experienced significantly lower cognitive load. However, despite this reduction, these students demonstrated lower-quality reasoning and argumentation in their final recommendations compared to those who used traditional search engines. Further, the homogeneity of the recommendations and justifications did not differ significantly between the two groups, suggesting that LLMs did not restrict the diversity of students' perspectives. These findings highlight the nuanced implications of digital tools on learning, suggesting that while LLMs can decrease the cognitive burden associated with information gathering during a learning task, they may not promote deeper engagement with content necessary for high-quality learning per se.

The vast and variable quality of online information demand critical digital skills (Vejvoda et al., 2023), especially when confronting contentious scientific topics. Whether information comes from real-world encounters, social media feeds, the news on TV, or search engines such as Google, we are constantly faced with the challenge of evaluating its accuracy (Simone et al., 2022; Vejvoda et al., 2023). Because of the ease, immediacy, and success with which one can obtain information, searching the web has become a daily routine to gain knowledge on a variety of topics, ranging from food safety (Bouzembrak et al., 2019) to issues of science (e.g., Lu & Reis, 2021). A large majority of Americans (81%) report they rely on their own web research over friends and family (43%) or professional experts (31%) when gathering information before making an important decision (Rainie et al., 2019).

Recent advances in artificial intelligence (AI), specifically in large language models (LLMs), are adding a new way learners may obtain information. ChatGPT, an LLM chatbot with vast knowledge about an array of topics, was released on November 30, 2022, and by February 2023, Microsoft and Google announced the upcoming availability of LLMs as well as Microsoft ending its waitlist for Bing Chat on May 4, 2023 (Spatharioti et al., 2023). Unlike traditional search engines, which direct users to websites based on their requests, LLMs aim to answer user questions directly based on a large amount of accumulated information.

In an extension of an approach suggested by Kammerer et al. (2021), who investigated university students' use of search engines to research an unsettled and unfamiliar socio-scientific issue (the use of nano-particles in sunscreen), this study investigates whether there are differences in the recommendations and justifications students reach when researching the same issue using search engines or LLMs. Based on cognitive load theory (Sweller, 2011) and on self-regulated learning (Zimmerman, 2000), we investigate the cognitive effort both types of information gathering require during a learning task as well as their recommendations and justifications.

1. Web search as learning

As digital technologies continue to permeate educational environments, the Internet has become an indispensable resource during learning. It is crucial that students harness this tool effectively, learning

* Corresponding author. LMU Klinikum, Institut für Didaktik und Ausbildungsforschung in der Medizin, Pettenkoferstraße 8a, 80336, München, Germany.
E-mail addresses: Matthias.Stadler@med.uni-muenchen.de (M. Stadler), Maria.Bannert@tum.de (M. Bannert), Michael.Sailer@uni-a.de (M. Sailer).

https://doi.org/10.1016/j.chb.2024.108386
Received 6 June 2024; Received in revised form 10 July 2024; Accepted 25 July 2024
Available online 30 July 2024
0747-5632/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

to maximize the potential of the information available to them. This requires more than mere access to data; it demands a set of skills that enhance comprehension and foster deep learning. The concept of "Search as Learning" (von Hoyer et al., 2022) posits that the act of searching the web can be a deliberate educational endeavor, rather than a passive consumption of information.

Consider a typical web search task where a student needs to gather information on climate change effects on different ecosystems. This task is not just about finding the right data; it involves discerning relevant information, evaluating sources, and synthesizing data into a coherent understanding (Fischer et al., 2014). During this process, students must exhibit a high degree of self-regulation to manage potential distractions, such as unrelated websites or social media notifications (Whipp & Chiarelli, 2004).

Self-regulation in learning, as defined by Zimmerman (2000), is a complex interaction of personal, behavioral, and environmental factors. This triadic model emphasizes that effective learning is not just about having the behavioral skills to manage environmental challenges but also involves possessing the requisite knowledge and a strong sense of personal agency. These skills enable learners to engage in self-generated thoughts, feelings, and actions that are systematically directed toward the attainment of personal goals. Correspondingly, the difficulty of web-based search tasks during learning can vary significantly based on several internal and external factors. Prior knowledge in the domain plays a crucial role; students with a deeper initial understanding of a topic can navigate information more strategically (Bannert, 2007; Moos & Azevedo, 2008; Willoughby et al., 2009). Additionally, the complexity of the search tasks can pose varying cognitive demands (Barsky & Bar-Ilan, 2012Barsky & Bar-Ilan, 2012). Finally, complex searches require a high degree of cognitive flexibility, metacognitive skills, and a robust working memory capacity, particularly in terms of the ability to shift attention between different pieces of information (Dommes et al., 2011) in respect to one's learning goals (Azevedo, 2005; Opfermann et al., 2012)

2. Cognitive load during web searches

The interplay between cognitive resources and learning outcomes is central to Cognitive Load Theory (CLT), which posits that human working memory has a limited capacity crucial for learning (Sweller, 2011, 2024). According to CLT, learning activities typically demand a significant allocation of cognitive resources, and instructional design plays a pivotal role in either exacerbating or alleviating cognitive load. The theory delineates three types of cognitive load: extraneous (ECL), intrinsic (ICL), and germane (GCL), each influencing learning in distinct ways (e.g., Bannert, 2002; Sweller et al., 2019).

Extraneous cognitive load refers to the load imposed by the manner in which information is presented to learners. In the context of web searches, the need to discern relevant from irrelevant information exemplifies additional ECL, such as when students encounter seductive details that may be interesting but are irrelevant to the task at hand (Sundararajan & Adesope, 2020). Intrinsic cognitive load is directly tied to the complexity of the material itself. In web searches, high ICL can occur when websites do not present information straightforwardly, or when the information has high element interactivity—requiring the learner to understand and integrate various components simultaneously (Mutlu-Bayraktar et al., 2019). The amount of ICL experiences depends on the learners' prior knowledge that helps them to organize the elements (Chen et al., 2017). Germane cognitive load, on the other hand, involves the cognitive resources devoted to the learner's active processing and automation of schemas. In the case of web searches, as students actively collect and synthesize information from various sources, they engage in processes that enhance their understanding and contribute to knowledge construction, thus resulting in higher GCL (Klepsch et al., 2017; Paas & van Gog, 2006).

3. Search engines vs. LLMs

The rise of large language models (LLMs) such as OpenAI's GPT has introduced a new paradigm for information gathering during learning that differs significantly from traditional search engines. Unlike search engines that direct users to relevant websites based on their keyword-based queries, LLMs aim to provide direct answers by "understanding" the intent of questions in human language and drawing upon vast knowledge bases to answer them.

While there is the obvious advantage of being able to write questions in plain text, rather than keyword-based queries (White et al., 2015) and receiving a plain text answer rather than having to construct the answer from an array of more or less relevant search results, there are several technical issues with the use of LLMs such as the issue of currency. Search engines continuously index the live web, aiming to provide users with the most up-to-date information available online (using different algorithms, depending on the specific search engine). In contrast, LLMs represent a fixed knowledge base that is only updated periodically, meaning the information they provide may not reflect the latest developments (Bender et al., 2021). Notably, though, some LLMs are now able to supplement their knowledge base through restricted web-searches. Still, search engines may provide an advantage in researching rapidly evolving topics or breaking news. Another concern is the potential for LLMs to generate "hallucinated" content - information that appears plausible but cannot be verified in the cited sources (Xu et al., 2024). While search engines direct users to original sources, LLM-generated responses do not always provide the same level of transparency, making it challenging for users to assess the reliability and accuracy of the information. Finally, the personalized and conversational nature of LLM interactions may amplify confirmation bias, as the systems tailor responses to align with users' existing beliefs and preferences (Sharma et al., 2024). This could lead to the reinforcement of echo chambers and the narrowing of information exposure, in contrast to the more diverse results often provided by search engines (Krupp et al., 2023).

On the other hand, the possibility of asking questions in natural language and the opportunity to ask for further explanation or refinement may represent an important advantage, reducing the cognitive demands of information gathering.

Table 1
Descriptive statistics of measures (n = 91).

Variable	Range	M	SD	ICL	ECL	GCL	Quality
ICL	1–7	3.81	1.65	(0.78)			
ECL	1–7	3.49	1.67	0.43*	(0.70)		
GCL	1–7	3.99	1.59	0.56*	0.48*	(0.69)	
Quality of justifications	0–4[a]	1.55	1.00	0.26*	0.25*	0.38*	–
Prior knowledge	0–7[b]	2.96	1.52	−0.18	−0.10	−0.00	0.21

Note. ICL = Intrinsic cognitive load, ECL = Extrinsic cognitive load, GCL = germane cognitive load. a = maximum value is 7, b = maximum value is 8. Internal consistencies (McDonald's Omega) are given in the diagonal. * indicates statistically significant correlations.

4. This study

These differences between researching using search engines and researching using LLMS, should also result in differences in the student's cognitive load and learning outcome. In this study, we test these assumptions following an approach introduced by Kammerer and colleagues (Kammerer et al., 2021), in which students research the use of nano-particles in sunscreen (an unsettled socio-scientific issue) with the goal of providing recommendations with justifications to a friend. Extending the original approach, the present study investigates the differences in cognitive load as well as the nature of recommendations and justifications drawn by university students when researching using either traditional web search engines or the large language model ChatGPT (version 3.5). Specifically, we explore the following research questions (RQ).

RQ1. Do students exhibit differences in extraneous, intrinsic, and germane cognitive load when using web search engines compared to interacting with LLMs to research the given scientific topic?

RQ2. Is there a difference in the quality of justifications presented in the students' final conclusions when using web search engines versus interacting with LLMs?

RQ3. Are there differences in the homogeneity of the recommendations in the students' final conclusions when using web search engines versus interacting with LLMs?

Regarding RQ1, we expect the research using LLMs to limit the amount of irrelevant information to be discarded and the content to be presented in a more accessible way (with the possibility of asking targeted follow-up questions to clarify specific aspects). Correspondingly, the need to actively engage with the content should be higher for the web-searches, which might result in a more in-depth processing. Based on these considerations, we deduct three hypotheses.

H1.1. Students using LLMs will experience lower extraneous cognitive load than students using traditional search engines.

H1.2. Students using LLMs will experience lower intrinsic cognitive load than students using traditional search engines.

H1.3. Students using LLMs will experience lower germane cognitive load than students using traditional search engines.

Regarding RQ2, there are two possible mechanisms that could be deducted from the theoretical models described above. The theory of self-regulated learning (Zimmerman, 2000) suggests that the reduced complexity of the search task using LLMs should result in a better understanding of the content. On the other hand, cognitive load theory (Sweller, 2024) predicts that the lower germane load of using LLMs, through the reduced need to make inferences, result in a reduced processing and thus worse memorization of the content. Therefore, we expect to find a difference between the groups but cannot clearly determine a direction of the effect.

H2. There will be a difference in the quality of justifications presented in the students' final conclusions between students using LLMs and students using traditional search engines.

A synthesis of the two theoretical approaches is suggested by Seufert et al. (2024), who describe the relation between cognitive load and self-regulated learning (i.e., the application of learning strategies; Seufert, 2020) as a tradeoff between cognitive load and cognitive resources. For simple tasks, mental resources are high and cognitive load is low, so there is no need to regulate, whereas for difficult tasks, load is high and resources may be too low to regulate. Consequently, learners engage in self-regulated learning to the fullest extent when confronted with tasks of medium difficulty because these tasks, there is an optimal level of load with sufficient amounts of resources available. It is unlikely that an information search task, as the one in our study, is too difficult for a sample of educated adults. However, it may be experienced as too simple to actually require much self-regulation. More regulation (more strategic behavior) should lead to better processing of the content to be learned. Consequently, we hypothesize that the amount of germane load (i.e., the type of load related to the processing, construction and automation of schemas) will mediate the effect of the type of information search on the quality of justifications presented in the students' final conclusions. We posit that a more demanding learning task will lead to the more germane load invested, which in turn will lead to better learning (Wang et al., 2023). However, this analysis is dependent on the results of H2, which we can only test exploratively.

Finally, search engines will provide an array of different results, whereas most LLMs aim to provide definitive answers to questions (Krupp et al., 2023). We therefore expect the students' final recommendations to be more homogenous when using LLMs than when using search engines.

H3. Students using LLMs will show less variation in their recommendations than students using traditional search engines.

5. Methods

5.1. Sample

In total, 92 students from various academic programs at a prestigious German university participated in the study between April and May 2023. Due to potential prior knowledge on the effect of nanoparticles in sunscreen, students in medicine, pharmacy, and biology were excluded as participants from the outset. One participant did not follow the instructions, using both an LLM and a search engine for the research, and was therefore excluded from the study. Thus, the final sample size was 91 students, with 67 female and 24 male participants. The average age of participants was 22.3 years (SD = 4.11). Participants were informed at the start of the study that their participation was voluntary and that all data would be anonymized to be used solely for research purposes. For certain programs, such as psychology majors and minors, participation-credits, a part of their curriculum, were offered as compensation.

5.1.1. Manipulation of the independent variable

Students were randomly assigned to one of two groups using different tools of information search. The first group was assigned the "web search" condition (n = 47) and the second group the "LLM" condition (n = 44). All students worked on university computers in a large lab allowing for multiple students working individually at the same time. For the web search condition, computers were configured so that the homepage in the Mozilla Firefox web browser was set to the Google search engine (google.com), for the LLM condition it was set to the

Table 2
ANCOVAs Statistics for investigating RQ1 and RQ2.

Hypothesis - variable	$M_{Web\ search}$ (SD)	M_{LLM} (SD)	*F*	*p*	η^2
H1.1 - ECL	3.96 (1.55)	3.00 (1.67)	8.26	0.005	0.09
H1.2 - ICL	4.43 (1.60)	3.16 (1.46)	16.48	<0.001	0.15
H1.3 - GCL	4.79 (1.18)	3.14 (1.53)	33.06	<0.001	0.27
H2 – Quality of justifications	1.87 (1.10)	1.20 (0.77)	11.18	0.001	0.11

Note. ICL = Intrinsic cognitive load, ECL = Extrinsic cognitive load, GCL = germane cognitive load.

chatbot ChatGPT (chat.openai.com) at time running the LLM GPT3.5. The study's questionnaires were opened in a second tab. Since it could not be assumed that every participant was familiar with ChatGPT or had prior knowledge of interacting with the system, the chat was prepared so that initially the following prompt was entered: "Introduce yourself and describe how to interact with you." After this input, ChatGPT wrote an introduction about itself and explained how to use the chatbot.

5.2. Procedure and learning task

The participating students first answered a set of demographic questions before being introduced to the search task. Following the approach introduced by Kammerer et al. (2021), students were instructed to help a fictitious friend named Paul decide whether to use sunscreen with mineral nanoparticles (i.e., particles of zinc oxide and titanium dioxide) in the future. Paul mentioned three advantages: the particles reflect UV light, thus filtering UV radiation without chemical agents that transform radiation into heat and can cause allergies or hormonal side effects; there are no known harmful side effects from the particles; and the use of nanoparticles in sunscreen can achieve very high SPF levels that protect the skin. However, Paul expressed concerns that sunscreens with nanoparticles might pose health risks. The task was to research whether his concerns about health risks were confirmed or if his fear was unfounded. Students were informed that they had exactly 20 min to research this issue after which they would be asked to provide a written recommendation with justifications without any notes (webpages or conversations with ChatGPT). The full instruction can be found on the osf repository (https://osf.io/jpxyt). To make sure there was no confounding of the cognitive load experienced during the web search task and while writing the recommendation, the CL-questionnaire was provided right after the task and before the recommendation was written. Finally, students answered a set of questions assessing their prior knowledge on nano-technology.

5.3. Measures

5.3.1. Cognitive load

A scale by Klepsch et al. (2017) was used to assess students' cognitive load during the learning task. The questionnaire consists of 7 statements covering ICL (e.g., "This task was very complex."), ECL (e.g., "In this task, it is difficult to recognize the most important information."), and GCL (e.g., "In this task, it is difficult to connect the central content with each other") to which the students rate their agreement on a scale from 1 (Do not agree at all) to 7 (Fully agree). Klepsch et al. (2017) report good internal consistency for all three scales based on two samples of university students. The full scale covering the 3 types of CL (ICL, ECL, and GCL) can be found on the osf repository (https://osf.io/jpxyt).

5.3.2. Justifications

In line with Kammerer et al. (2021) and following previous research in the area of multiple document comprehension (Bråten et al., 2018), we operationalized the quality of the justifications by coding the students' recommendations and justifications in terms of whether or not they mentioned that (a) risks are low or no risks are known and/or that advantages predominate, (b) a coating procedure could reduce the potential risks associated with nanoparticles, (c) there is (only) risk when the skin is damaged, (d) sprays are risky because nanoparticles could be inhaled, (e) risks are high or uncertain, (f) the advantage is only cosmetic, and (g) mineral sunscreen without nanoparticles could be used instead. The coding scheme emerged from both a consultation of nanotechnology experts (from the Leibniz Institute for New Materials, Saarbrücken, Germany) on relevant aspects of the topic and a bottom-up analysis of the written justifications. We used the number of relevant aspects each participant mentioned in their recommendation as a dependent measure. Two raters independently coded all justifications, achieving an overall inter-rater agreement of $k = 0.92$. The disagreements were resolved through discussion. The coding scheme was not available to the participants.

Table 3
Frequency of students' recommendations regarding the use of nano-particles in sunscreen.

	Recommendation			Total
	Against	Neutral	In favor	
LLM	9	7	28	44
Web search	15	8	24	47
Total	24	15	52	91

5.3.3. Recommendation

Regarding the recommendation that participants make in their recommendations with justifications, answers were coded as recommending the use of sunscreen with nanoparticles (+1), neither recommending nor not advising against it (0), and advising against it (−1). Two raters independently coded all recommendations, achieving an overall inter-rater agreement of $k = 0.89$. The disagreements were resolved through discussion.

5.3.4. Prior domain knowledge

To measure students' prior knowledge regarding nanotechnology, we used an adapted version of the Public Knowledge of Nanotechnology (PKNT) test by Lin et al. (2013). On this knowledge test, participants answered eight multiple-choice questions with four response alternatives about different concepts in nanotechnology, such as size and scale, structure of matter, or current applications of nanomaterials. The quality indicator and the score for prior knowledge, as index-scores, were not expected to be internally consistent (Stadler et al., 2021). All items can be found on the osf repository (https://osf.io/jpxyt).

5.4. Statistical analysis

All analyses were conducted using jamovi 2.3 (The jamovi project, 2003). To detect potential prior knowledge differences between both groups, we conducted a *t*-test. To test the Hypotheses based on RQ1 and RQ2, we conducted analyses of covariance (ANCOVA) comparing the mean scores of the two groups controlling for prior knowledge. For the exploratory mediation analysis in extension of H2, we defined a mediation model using the jAMM module (Gallucci, 2020). The model defined direct effects of the group, GCL and prior knowledge on the quality of the justifications with an additional indirect effect of GCL mediating the effect of the group on the quality of the justifications. H3, differences in the variation in students' recommendations, was tested using a Chi-Squared test. Internal consistency was estimated for the cognitive load scales using McDonald's Omega. All analyses were conducted using jamovi 2.3 (Şahin & Aybek, 2020). The alpha level was set to 5%.

6. Results

6.1. Descriptive statistics

All descriptive statistics for the continuous dependent variables and prior knowledge are provided in Table 1. Notably, the average quality of the justifications was rather low with a maximum of 4 out of 7 possible points reached. The internal consistency of the cognitive load scales was acceptable given the low number of items. There was substantial intercorrelation between the different facets of cognitive load. The quality indicator showed small correlations to all of the facets of cognitive load. Students had some prior knowledge, which was not significantly related to any of the dependent variables.

6.1.1. Inferential statistics

First, we found in the randomization check that the two groups did not differ significantly in their level of prior knowledge ($t(89) = -0.28$; $p = 0.777$, $d = -0.06$).

In line with Hypotheses H1.1 to H1.3, students using LLMs experienced substantially lower cognitive load than students using traditional search engines on all three facets of cognitive load (Table 2) adjusted for prior knowledge. The strongest difference was found for GCL, the smallest for ECL.

Also, in line with H2, there was a significant difference in the quality of justifications presented in the students' final conclusions between students using LLMs and students using traditional search engines (Table 2) adjusted for prior knowledge. Students in the web search condition listed significantly more relevant arguments in their statements than students in the LLM condition.

In an exploratory extension of H2, we tested, whether the difference in the quality of the arguments and reasoning presented in the students' final conclusions between students using LLMs and students using traditional search engines was mediated by the difference in GCL. We found a significant indirect effect ($\beta = 0.15$; $p = 0.020$) and no significant direct effect ($\beta = 0.19$; $p = 0.095$) indicating a full mediation of the effect of the different research conditions on the quality of the arguments and reasoning presented through GCL.

Contrary to H3, students using LLMs did not show less variation in their recommendations than students using traditional search engines ($\chi^2(2) = 1.78$; $p = 0.411$). In fact, the distribution of recommendations was similar for the two groups (Table 3).

7. Discussion

The results of the current study offer several intriguing insights into the differences in cognitive load and the quality of learning outcomes between traditional web searches and those conducted using LLMs such as ChatGPT. This investigation builds on previous research by exploring how these different information-seeking approaches affect the learning process (e.g., Spatharioti et al., 2023), specifically within the context of a complex socio-scientific issue like the use of nanoparticles in sunscreen (Kammerer et al., 2021).

The findings suggest significant differences in cognitive load between the two groups, with students using LLMs experiencing lower cognitive loads across the extraneous, intrinsic, and germane facets. This supports the hypotheses (H1.1 and H1.2) that LLMs, by providing direct, concise answers, may reduce the cognitive burden associated with sifting through and synthesizing multiple web sources. In line with cognitive load theory, this reduction in cognitive load could be beneficial for learning by freeing up cognitive resources (Sweller, 2011, 2024). However, in line with our hypothesis H1.3, the lower GCL in the LLM group suggests that while the information was easier to process, it might not have engaged the deep learning processes as effectively as the more challenging traditional search tasks.

Interestingly, despite the reduced ICL and ECL, students in the LLM condition presented justifications of lower quality than those in the traditional search engine condition (H2). This outcome supports the findings of prior studies suggesting that a higher degree of interaction with diverse and sometimes challenging information—as often encountered in traditional searches—may promote better comprehension and processing of learning material (e.g., Cierniak et al., 2009). This interaction likely encourages students to engage more deeply with the content, thus enhancing their learning experience and leading to more sophisticated justifications in their conclusions. The results of our exploratory mediation analysis, suggesting a complete mediation of the effect of the two research conditions through GCL, further support this interpretation. These results are also in line with the model proposed by Seufert (2020). According to this model, our results would suggest that research using LLMs is not cognitively challenging enough, that students regulate their learning behavior and engage the content systematically. On the other hand, researching using the search engine was not too challenging so that students could still process the content provided systematically (Seufert et al., 2024).

Contrary to our third hypothesis (H3), there was no significant difference in the homogeneity of recommendations between the two groups. This suggests that despite the more structured and directed responses provided by LLMs, students still reached a diverse set of conclusions. This finding contradicts concerns that LLMs might lead to a narrowing of perspectives due to their design to provide seemingly definitive answers (Krupp et al., 2023). It appears that even within the constraints of LLM responses, there is room for interpretation and individual judgment among users.

These results collectively underscore the complex relationship between the method of information presentation and learning outcomes. While LLMs can reduce the cognitive load, which is theoretically beneficial for learning, the nature of the content delivered and the interaction required to engage with that content also play crucial roles. The ease of obtaining answers from LLMs does not necessarily translate to deeper learning or better-quality outcomes, as evidenced by the superior performance of the Web search group in developing more detailed and higher-quality justifications. Specifically, the results point towards the importance of inferring cognitive processes (e.g., reorganizing, reflecting; Chi & Wylie, 2014) when using digital technologies that are associated with higher learning outcomes (Sailer et al., 2024). While recall of knowledge from LLMs might more closely be related to shallow learning processes, inferring knowledge from different sources might more closely be related to deep learning processes.

Summing up, this study also emphasizes the importance of active engagement with content for deeper learning, as suggested by the cognitive load theory as well as the self-regulated learning framework. The findings highlight the need for educational strategies that not only provide information efficiently but also challenge learners to engage actively with complex material.

7.1. Limitations and future directions

This study, while informative, has several limitations. First, the absence of a "think aloud" protocol, as used in Kammerer et al. (2021), limits the insights into the cognitive and metacognitve processes students engaged in during their searches (Bannert & Mengelkamp, 2008). Future research could incorporate such qualitative measures to gain a deeper understanding of how students interact with and process information from different sources. Additionally, examining the search logs could provide insights into the strategies used by students, which would help further explain the different learning trajectories observed (Fan et al., 2022; Winne, 2023). Effective prompting strategies fundamentally change the utility of LLMs and factors such as previous experience with LLMs as well as knowledge on prompting could be important moderators that need to be explored further (Knoth et al., 2024). This could be an important direction for future studies, helping to refine our understanding of how digital tools influence learning. Moreover, a closer inspection of the participants' use of their test time might reveal whether the allotted time was too short or even too long. Especially the participants using the LLMs might have required much less time to complete the task but were forced to spend a total of 20 min on it.

Another significant limitation concerns the generalizability of our results. The study sample was rather small, not representatively sampled and consisted solely of university students who are likely well-versed in web searches and possess high reading ability. While the sample size was sufficient to detect effects of practical relevance with acceptable power, this particular demographic does not represent the broader population (e.g., Kammerer et al., 2015), which varies widely in both digital literacy and cognitive skills (Hahnel et al., 2018). Future studies should consider a more diverse sample to determine if the findings observed in this research extend to populations with different educational backgrounds, varying levels of familiarity with technology, and diverse cognitive

capabilities. This expansion of the participant pool would provide a more comprehensive understanding of the impacts of LLMs and traditional web searches across different societal segments. Furthermore, real-world learning would likely integrate a multitude of information sources, rather than being artificially constrained to LLMs or search engines. In particular, since the completion of this study, numerous search engines have incorporated LLMs to enhance their comprehension of user queries, while LLMs have also gained the capacity to perform limited web searches. Future research should investigate the orchestration of different digital tools and resulting learning strategies.

A final minor limitation concerns the timing of the knowledge test. The test was purposefully conducted after the search task. This ensured that the questions of the knowledge test would not influence the students' searching by guiding them towards relevant keywords. On the other hand, it is possible that engaging in the search task could have increased students' knowledge on nano-particles, thus inflating their test scores. Comparing these two risks, the former (knowledge test after the search task) seemed like the smaller limitation to the interpretability of our results than the latter.

8. Conclusion

In conclusion, while LLMs like ChatGPT offer an efficient way to reduce intrinsic and extraneous cognitive load, they may not always facilitate the deep learning necessary for complex decision-making tasks. Traditional search engines, by necessitating more active engagement, may promote a higher quality of learning, underscoring the need for educational practices that encourage critical engagement with diverse information sources.

CRediT authorship contribution statement

Matthias Stadler: Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria Bannert:** Writing – review & editing, Validation, Conceptualization. **Michael Sailer:** Writing – original draft, Validation, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is available on osf

References

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199–209. https://doi.org/10.1207/s15326985ep4004_2

Bannert, M. (2002). Managing cognitive load—recent trends in cognitive load theory. *Learning and Instruction, 12*(1), 139–146. https://doi.org/10.1016/S0959-4752(01)00021-4

Bannert, M. (2007). *Metakognition beim Lernen mit Hypermedien*. Waxmann Verlag.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*(1), 39–58. https://doi.org/10.1007/s11409-007-9009-6

Barsky, E., & Bar-Ilan, J. (2012). The impact of task phrasing on the choice of search keywords and on the search process and success. *Journal of the American Society for Information Science and Technology, 63*(10), 1987–2005. https://doi.org/10.1002/asi.22654

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. In *ACM digital library, proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bouzembrak, Y., Klüche, M., Gavai, A., & Marvin, H. J. (2019). Internet of Things in food safety: Literature review and a bibliometric analysis. *Trends in Food Science & Technology, 94*, 54–64. https://doi.org/10.1016/j.tifs.2019.11.002

Bråten, I., Brante, E. W., & Strømsø, H. I. (2018). What really matters: The role of behavioural engagement in multiple document literacy tasks. *Journal of Research in Reading, 41*(4), 680–699. https://doi.org/10.1111/1467-9817.12247

Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review, 29*(2), 393–405. https://doi.org/10.1007/s10648-016-9359-1

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Computers in Human Behavior, 25*(2), 315–324. https://doi.org/10.1016/j.chb.2008.12.020

Dommes, A., Chevalier, A., & Lia, S. (2011). The role of cognitive flexibility and vocabulary abilities of younger and older users in searching for information on the web. *Applied Cognitive Psychology, 25*(5), 717–726. https://doi.org/10.1002/acp.1743

Fan, Y., van der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., Moore, J., Molenaar, I., Bannert, M., & Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. *Metacognition and Learning, 17*(3), 949–987. https://doi.org/10.1007/s11409-022-09291-1

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research, 2*(3), 28–45. https://eric.ed.gov/?id=ej1090940.

Gallucci, M. (2020). jAMM: Jamovi advanced mediation models [Computer software] https://jamovi-amm.github.io/.

Hahnel, C., Goldhammer, F., Kröhne, U., & Naumann, J. (2018). The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior, 78*, 223–234. https://doi.org/10.1016/j.chb.2017.10.004

Kammerer, Y., Amann, D. G., & Gerjets, P. (2015). When adults without university education search the Internet for health information: The roles of Internet-specific epistemic beliefs and a source evaluation intervention. *Computers in Human Behavior, 48*, 297–309. https://doi.org/10.1016/j.chb.2015.01.045

Kammerer, Y., Gottschling, S., & Bråten, I. (2021). The role of internet-specific justification beliefs in source evaluation and corroboration during web search on an unsettled socio-scientific issue. *Journal of Educational Computing Research, 59*(2), 342–378. https://doi.org/10.1177/0735633120952731

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2023). Unreflected acceptance – investigating the negative consequences of ChatGPT-assisted problem solving in physics education. August 21 http://arxiv.org/pdf/2309.03087v1.

Lin, S.-F., Lin, H., & Wu, Y. (2013). Validation and exploration of instruments for assessing public knowledge of and attitudes toward nanotechnology. *Journal of Science Education and Technology, 22*(4), 548–559. https://doi.org/10.1007/s10956-012-9413-9

Lu, T., & Reis, B. Y. (2021). Internet search patterns reveal clinical course of COVID-19 disease progression and pandemic spread across 32 countries. *NPJ Digital Medicine, 4*(1), 22. https://doi.org/10.1038/s41746-021-00396-6

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*(2), 270–298. https://doi.org/10.1016/j.cedpsych.2007.03.001

Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education, 141*, Article 103618. https://doi.org/10.1016/j.compedu.2019.103618

Opfermann, M., Azevedo, R., & Leutner, D. (2012). Metacognition and hypermedia learning: How do they relate? In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 2224–2228). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_709.

Paas, F., & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16*(2), 87–91. https://doi.org/10.1016/j.learninstruc.2006.02.004

Rainie, L., Keeter, S., & Perrin, A. (2019). Trust and distrust in America: Many Americans think declining trust in the government and in each other makes it harder to solve key problems. *They have a wealth of ideas about what's gone wrong and how to fix it.* Pew Research Center.

Şahin, M., & Aybek, E. (2020). Jamovi: An easy to use statistical software for the social scientists. *International Journal of Assessment Tools in Education, 6*(4), 670–692. https://doi.org/10.21449/ijate.661803

Sailer, M., Maier, R., Berger, S., Kastorff, T., & Stegmann, K. (2024). Learning activities in technology-enhanced learning: A systematic review of meta-analyses and second-order meta-analysis in higher education. *Learning and Individual Differences, 112*, Article 102446. https://doi.org/10.1016/j.lindif.2024.102446

Seufert, T. (2020). Building bridges between self-regulation and cognitive load—an invitation for a broad and differentiated attempt. *Educational Psychology Review, 32*(4), 1151–1162. https://doi.org/10.1007/s10648-020-09574-6

Seufert, T., Hamm, V., Vogt, A., & Riemer, V. (2024). The interplay of cognitive load, learners' resources and self-regulation. *Educational Psychology Review, 36*(2). https://doi.org/10.1007/s10648-024-09890-1

Sharma, N., Liao, Q. V., & Xiao, Z. (2024). Generative echo chamber? Effects of LLM-powered search systems on diverse information seeking. February 8 http://arxiv.org/pdf/2402.05880v2.

Simone, C. de, Battisti, A., & Ruggeri, A. (2022). Differential impact of web habits and active navigation on adolescents' online learning. *Computers in Human Behavior Reports, 8*, Article 100246. https://doi.org/10.1016/j.chbr.2022.100246

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment. July 7 http://arxiv.org/pdf/2307.03744v2.

Stadler, M., Sailer, M., & Fischer, F. (2021). Knowledge as a formative construct: A good alpha is not always better. *New Ideas in Psychology, 60*, Article 100832. https://doi.org/10.1016/j.newideapsych.2020.100832

Sundararajan, N., & Adesope, O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review, 32*(3), 707–734. https://doi.org/10.1007/s10648-020-09522-4

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (pp. 37–76). Elsevier. https://doi.org/10.1016/b978-0-12-387691-1.00002-8.

Sweller, J. (2024). Cognitive load theory and individual differences. *Learning and Individual Differences, 110*, Article 102423. https://doi.org/10.1016/j.lindif.2024.102423

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 Years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

The jamovi project. (2003). jamovi [Computer software] Version 2.3. https://www.jamovi.com.

Vejvoda, J., Stadler, M., Schultz-Pernice, F., Fischer, F., & Sailer, M. (2023). Getting ready for teaching with digital technologies: Scenario-based self-assessment in teacher education and professional development. *Unterrichtswissenschaft, 51*(4), 511–532. https://doi.org/10.1007/s42010-023-00186-x

von Hoyer, J., Hoppe, A., Kammerer, Y., Otto, C., Pardi, G., Rokicki, M., Yu, R., Dietze, S., Ewerth, R., & Holtz, P. (2022). The search as learning spaceship: Toward a comprehensive model of psychological and technological facets of search as learning. *Frontiers in Psychology, 13*, Article 827748. https://doi.org/10.3389/fpsyg.2022.827748

Wang, T., Li, S., Huang, X., Pan, Z., & Lajoie, S. P. (2023). Examining students' cognitive load in the context of self-regulated learning with an intelligent tutoring system. *Education and Information Technologies, 28*(5), 5697–5715. https://doi.org/10.1007/s10639-022-11357-1

Whipp, J. L., & Chiarelli, S. (2004). Self-regulation in a web-based course: A case study. *Educational Technology Research & Development, 52*(4), 5–21. https://doi.org/10.1007/BF02504714

White, R. W., Richardson, M., & Yih, W. (2015). Questions vs. Queries in informational search tasks. In A. Gangemi, S. Leonardi, & A. Panconesi (Eds.), *Proceedings of the 24th international conference on world wide web* (pp. 135–136). ACM. https://doi.org/10.1145/2740908.2742769.

Willoughby, T., Anderson, S. A., Wood, E., Mueller, J., & Ross, C. (2009). Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge. *Computers & Education, 52*(3), 640–648. https://doi.org/10.1016/j.compedu.2008.11.009

Winne, P. H. (2023). Roles for information in trace data used to model self-regulated learning. In V. Kovanovic, R. Azevedo, D. C. Gibson, & D. Ifenthaler (Eds.), *Advances in analytics for learning and teaching. Unobtrusive observations of learning in digital environments* (pp. 175–196). Springer International Publishing. https://doi.org/10.1007/978-3-031-30992-2_11.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. January 22 http://arxiv.org/pdf/2401.11817v1.

Zimmerman, B. J. (2000). Attaining self-regulation. In *Handbook of self-regulation* (pp. 13–39). Elsevier. https://doi.org/10.1016/B978-012109890-2/50031-7.