PNAS NEXUS

PNAS Nexus, 2025, 4, pgaf316
https://doi.org/10.1093/pnasnexus/pgaf316
Advance access publication 28 October 2025
Research Report

Experimental evidence of the effects of large language models versus web search on depth of learning

Shiri Melumad [a,*] and Jin Ho Yun [b,c]

[a]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
[b]College of Business, New Mexico State University, Las Cruces, NM 88003, USA
[c]Wharton Neuroscience Initiative, University of Pennsylvania, Philadelphia, PA 19104, USA
*To whom correspondence should be addressed: Email: melumad@wharton.upenn.edu
Edited By David Rand

Abstract

The effects of using large language models (LLMs) versus traditional web search on depth of learning are explored. A theory is proposed that when individuals learn about a topic from LLM syntheses, they risk developing shallower knowledge than when they learn through standard web search, even when the core facts in the results are the same. This shallower knowledge accrues from an inherent feature of LLMs—the presentation of results as summaries of vast arrays of information rather than individual search links—which inhibits users from actively discovering and synthesizing information sources themselves, as in traditional web search. Thus, when subsequently forming advice on the topic based on their search, those who learn from LLM syntheses (vs. traditional web links) feel less invested in forming their advice, and, more importantly, create advice that is sparser, less original, and ultimately less likely to be adopted by recipients. Results from seven online and laboratory experiments ($n = 10{,}462$) lend support for these predictions, and confirm, for example, that participants reported developing shallower knowledge from LLM summaries even when the results were augmented by real-time web links. Implications of the findings for recent research on the benefits and risks of LLMs, as well as limitations of the work, are discussed.

Keywords: large language models, web search, search process, learning, natural language processing

Significance Statement

Might the ease afforded by large language model (LLM) syntheses come at the cost of learning compared with traditional web search? A theory is proposed that because LLM summaries lessen the need to discover and synthesize information from original sources—steps essential for deep learning—users may develop shallower knowledge compared with learning from web links. When subsequently forming advice on the topic, this manifests in advice that is sparser, less original—and less likely to be adopted by recipients. Results from seven experiments support these predictions, showing that these differences arise even when LLM summaries are augmented by real-time web links, for example. Hence, learning from LLM syntheses (vs. web links) can, at times, limit the development of deeper, more original knowledge.

Introduction

Since their public release in 2022, large language models (LLMs) such as ChatGPT have become an increasingly pervasive tool for acquiring information (1–4). A defining feature of LLMs is the format in which their results are displayed: whereas traditional online search presents a series of web links that users must navigate and distill on their own, LLMs complete this process on behalf of users, providing automatic syntheses of vast amounts of information (5, 6). Hence, LLMs enable users to learn about topics faster and with less effort than in traditional web search (6, 7). It is perhaps unsurprising, then, that major search engines such as Google now offer "AI Overviews" of their standard search results, putting LLMs at the forefront of search while making web links a secondary resource (5).

While LLMs afford obvious efficiency gains for users, we propose that this greater ease may come at a cost: that of reducing the depth of knowledge, and originality of thought, gleaned from one's search in certain contexts. Specifically, while the need to navigate and summarize different information sources might make learning through web search more effortful, it can offer the often-overlooked benefit of constructing deeper, and more unique, knowledge structures (8, 9)—something that is less likely when various information sources are already synthesized for the user by an LLM. As a result, individuals who learn about a topic from LLM syntheses (vs. standard web links) may at times emerge feeling that they have learned less about it, and, in turn, generate ideas about it that are shallower—for example, by forming advice for others on the subject that is sparser, more

Competing Interest: The authors declare no competing interests.
Received: March 28, 2025. **Accepted:** September 11, 2025

generic, and, ultimately, less informative and persuasive to others.

We tested these ideas across a series of experiments in which participants were randomly assigned to learn about a topic either from standard Google web links or LLM syntheses (e.g. ChatGPT) and were then asked to create advice on the subject based on what they learned. Results from seven online and laboratory experiments ($n = 10{,}426$) yield robust evidence that in this task context, participants reported developing shallower knowledge when learning about a topic from LLM summaries (vs. web links), and that this occurred because those who used an LLM exerted less effort in learning from its synthesized responses compared with those who gathered and distilled information themselves through web links. As a result, when they subsequently wrote advice on the subject, those who learned through an LLM (vs. web search) were not only less invested in forming their advice, but they also created content that was objectively sparser and more generic—such that recipients found it to be less informative and were less likely to adopt it. These results were robust across different search topics and LLM tools and held even when, for example, LLM summaries were augmented by real-time web links.

Theoretical background

LLMs as aids to humans

Reflecting their widespread and rapid adoption, a large literature has emerged on the usefulness of LLMs in aiding human decision-making. A central finding of this work is that LLMs can be highly effective at solving a wide range of human reasoning problems, both alone and in collaboration with humans. For example, Kung et al. (10) found that GPT-3.5 could pass the US Medical Licensing Exam, and Terwiesch et al. (11) reported that GPT-3 could answer MBA exam questions with an accuracy that equaled or exceeded that of MBA students. Likewise, when LLMs are used to assist humans, they can improve problem solving compared with humans alone. For example, allowing professional consultants to use GPT-4 was shown to increase the speed and accuracy with which business analyses were solved (12), and use of ChatGPT increased the quality and speed with which professional writing tasks were completed (13). Additionally, use of Claude-3.5-Sonnet and ChatGPT-3.5 as aids in creativity tasks improved the average novelty of ideas compared with users who did not use LLMs (14, 15).

The evidence on the value of LLMs as an aid to learning, however, has been more equivocal. Most of the extant evidence comes from comparisons of LLM-aided versus unaided (human-led) student achievement, where several studies have found that using LLMs can improve academic performance in tasks such as coding (16) and second-language writing (17, 18). Conversely, however, other work has found that academic performance can be harmed in tasks where students need to generalize from the information they initially acquired from the LLM (19, 20). For example, Bastani et al. (19) found that high school students who relied on a GPT-4 learning aid to complete practice math assignments subsequently performed worse on new problems once they were deprived of the aid, compared with those who had never accessed it.

Why might the use of LLMs impede learning in certain contexts? Although the mechanism has not been fully resolved, some researchers have suggested that this might occur because students excessively rely on LLMs as "crutches" when developing knowledge (19, 21). That is, the availability of "quick and easy" answers from LLMs may prevent students from developing the skills needed to solve problems in new contexts on their own.

In this work, we examine how one's ability to learn about a topic is affected by the use of LLMs relative to a more traditional tool for information acquisition: standard web search. While some recent work has compared the use of LLMs versus Google search (7, 22–24), these depart from our effort by focusing on the relative accuracy and efficiency with which users retrieve information (7, 23, 24). Here, we explore the distinct question of how a definitional feature of LLMs—the ability to rapidly synthesize vast arrays of information—can fundamentally alter the depth of knowledge users may develop from their search. Specifically, whereas traditional web search requires users to expend effort in navigating web links, reading different sources, and interpreting and synthesizing information, LLMs carry out this distillation process for users. Thus, while LLMs offer clear efficiency benefits compared with traditional web search, we suggest that this ease can come at the cost of deeper knowledge development in certain contexts. We elaborate on these ideas in turn.

Learning through web search versus LLMs

A central thesis of this work is that whereas LLMs provide a faster route to finding answers than web search, by doing so they inhibit a process that can be instrumental to learning: the self-guided exploration of different information that requires original synthesis. The value of self-directed knowledge acquisition for skill development has been explored in the literature on "search-as-learning," which argues that when we engage in traditional web search on platforms like Google, we do more than accumulate facts—we also develop knowledge structures through an iterative process of posing queries, gathering and interpreting information from different websites, and then assembling this knowledge into a cohesive whole (8, 9, 25, 26). Thus, the process of "sensemaking" through web search can be highly dynamic for users, marked by the recursive process of synthesizing and revising (27, 28). In contrast, we argue that since LLMs are designed to perform such sensemaking on behalf of the user, this critical ingredient in learning is often diminished relative to gathering information from web links.

The notion that the ease afforded by LLMs (vs. web search) might suppress depth of learning is also broadly consistent with past work on "desirable difficulty" in other domains. This literature reports that making learning easier does not always improve learning outcomes (29, 30). As an example, Cockburn et al. (31) focused on how to improve people's ability to learn the location of graphical objects (e.g. icons) on user interfaces. They found that when learning was made more effortful—there, by requiring users to brush aside a "frost" that obscured each object to reveal its identity—participants subsequently had better memory for locations compared with those who did not need to un-obstruct their view of the object. Likewise, Alter et al. (32) found that participants performed better on a cognitive reflection test when the font in the task was made smaller and more difficult to read, and Diemand-Yauman et al. (33) similarly found that students earned better grades when the font on worksheets was harder to read.

Perhaps the most widely accepted explanation for why task disfluency can at times lead to better learning outcomes is that when we encounter friction in learning, we reflexively devote more cognitive resources to overcoming it, which leads us to process what we are trying to learn more deeply (30). In the context studied here, learning through web search involves a particular type of disfluency: the need to engage in trial-and-error navigation among result links, discovering which are suitable for the question at hand, and then interpreting and synthesizing the different

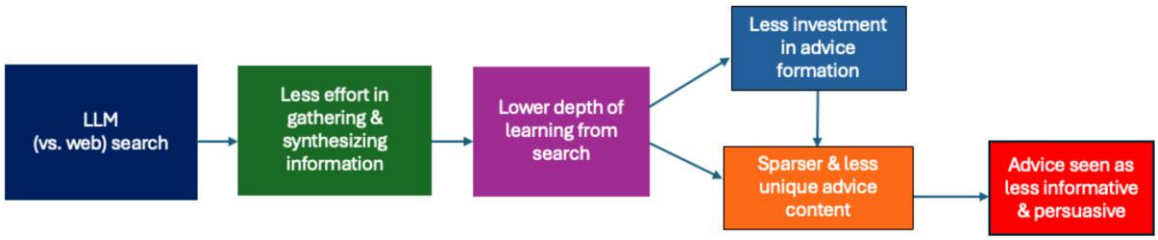

Fig. 1. Proposed conceptual model.

pieces of information for oneself (9, 26). Hence, we expect that the effort required to learn from standard web search results can lead users to develop deeper knowledge on a subject.

In contrast to web search, when learning from LLM summaries users no longer need to exert the effort of gathering and distilling different informational sources on their own—the LLM does much of this for them. We predict that this lower effort in assembling knowledge from LLM syntheses (vs. web links) risks suppressing the depth of knowledge that users gain, which subsequently affects the nature of the advice they form on the topic for others. Specifically, after learning about a subject via LLM summaries (vs. web links), users will be less invested in writing the advice and, more importantly, their advice will reflect shallower knowledge, such as by being sparser and more generic (34, 35). As a result, recipients of this advice will find the advice to be less informative and will ultimately be less willing to adopt it. We illustrate these ideas in the conceptual model in Fig. 1.

Empirical analysis

We report the results of seven total experiments that test these predictions, including four experiments in the main text and three in the Supplementary Material. The first preregistered experiment provides an initial test of our predictions in a naturalistic setting where participants are asked to learn about a particular topic by engaging interactively with either Google or ChatGPT, and then to provide advice to a friend based on what they learned from their search. In the second preregistered experiment, we report a more conservative test of our predictions by exposing participants to search results containing the *same* set of facts, varying only whether it is presented in the format of an LLM summary or a set of linked websites. In the third experiment, we test for the robustness of the effects in a laboratory setting, this time by holding constant the search engine—Google search—and varying whether participants learned about a topic through standard Google search results or Google's LLM synthesis ("AI Overview") presented at the top of the standard search results page. Finally, in the fourth experiment, we explore the downstream consequences of these effects by presenting advice written by participants in a prior experiment to an independent set of "recipients"—blind to the original search platform used to learn about the topic—and examining their willingness to adopt the advice. In the Supplementary Material, we report the results of three preregistered replications that, for example, test for the robustness of the basic effects to the inclusion of real-time web links in LLM syntheses and to a search topic that was of high personal relevance.

All experimental procedures involving human participants were approved by the University of Pennsylvania Institutional Review Board (IRB Protocol #854440). Informed consent was obtained from all participants prior to their participation in the study, in accordance with the university's human subjects protection guidelines.

Experiment 1: testing for the effects of ChatGPT versus Google search use

Methods

Participants were 1,136 members[a] of the Prolific panel who were randomly assigned to learn about a topic using either ChatGPT or Google and then to write advice for their friend on it (M_{age} = 42.40, SD = 13.46; 46% female, 53% male, 1% nonbinary). For the first experiment, we built an in-house platform that allowed participants to conduct actual ChatGPT or Google searches while enabling us to record their submitted queries. Per the preregistered criteria, we excluded 32 participants for failing attention checks, resulting in a final sample of 1,104 participants (M_{age} = 42.40, 53% female). The preregistration for the study is available at https://aspredicted.org/D1G_BRC.

In both conditions, participants were first asked to imagine that a friend was seeking advice about how to plant a vegetable garden, and that they wanted to search for information about the topic before forming their advice on it. Participants in the Google condition were then redirected to a website containing the Google search interface and, after typing and submitting their query, they were shown the actual Google search results that it yielded (see Fig. 2A for an illustrative screenshot).[b] The Google search results were presented across different pages, with each page containing 10 search links. Participants were instructed to browse through as many linked websites as they wanted to, and they could restart their search and submit a new query as many times as they wished. A timer was embedded within the website to record the amount of time participants spent on the search task, which included the time spent generating queries and engaging with the search results. This served as a proxy measure for the amount of effort they put into learning from the results.

In the ChatGPT condition, participants submitted their search prompts to a ChatGPT interface that was programmed into the Qualtrics survey (see Fig. 2B for illustrative screenshot).[c] Similar to the Google condition, participants could interactively engage with the in-house ChatGPT, inputting as many prompts as they wished, and a timer was embedded in the survey to provide a proxy measure for the amount of effort invested in learning from the results.

Once participants finished the search task, they proceeded to write their advice on the topic within the survey. After writing their advice, they rated the extent to which they agreed with a series of items that captured the focal measures (on a 1: "Strongly disagree" to 5: "Strongly agree" scale). To capture the degree to which participants felt they learned from the search task—the focal construct of interest—they rated their agreement with three items: "I learned new things on the topic from the Google/ChatGPT results," "The Google/ChatGPT results provided comprehensive information on the subject," and "I feel a sense of personal ownership over what I learned." To capture how invested they were in forming their advice, participants rated their agreement

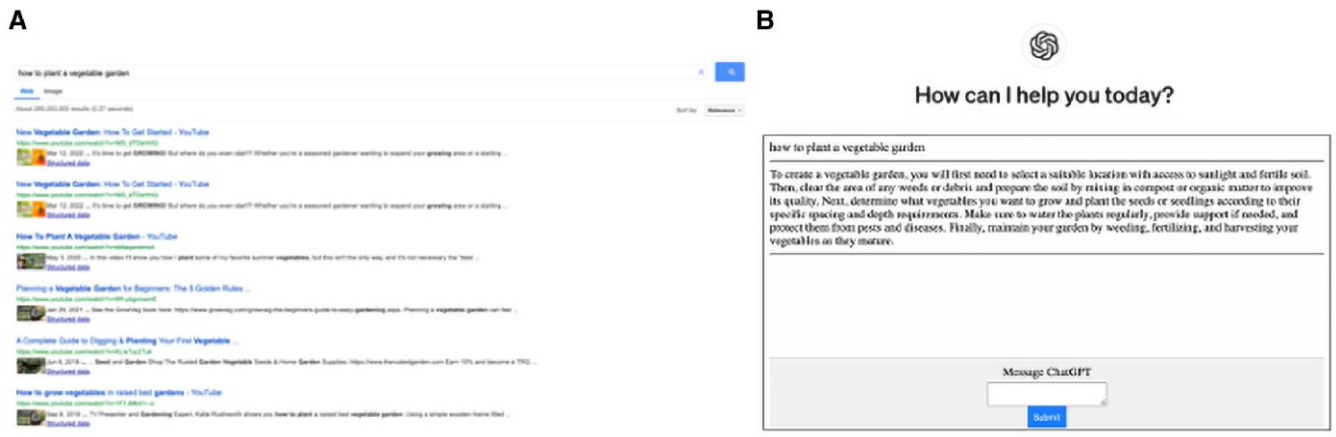

Fig. 2. Screenshots of the (A) Google search and (B) ChatGPT interface of experiment 1.

with two items: "I put a lot of thought and effort into generating my advice" and "I feel a sense of personal ownership over the advice I gave." To provide a third, proxy measure of how invested participants were in forming the advice, a timer was embedded into the Qualtrics page where participants completed the advice task, allowing us to record the amount of time they spent writing it. The full survey instrument is reported in the Supplementary Material.

Once data collection was completed, a battery of natural language processing (NLP) tools was used to test the prediction that advice written after learning from ChatGPT (vs. Google) would contain linguistic indicators of shallower (vs. deeper) knowledge. Past research on knowledge development contends that deeper learning of a topic tends to be marked by a greater retention of relevant facts, and enhanced ability to elaborate on those facts in an original manner (34, 35). Building on this, in the context studied here, deeper (vs. shallower) knowledge would manifest in advice content containing more references to facts, higher word count, and language that is more unique to the writer (34, 35). To capture the degree of references to factual entities in participants' advice, we used the pretrained Named Entity Recognition tool within the Python spaCy NLP library (36), which calculates the number of individual words or word pairs in a corpus that are associated with each of seven predefined topical categories, such as products, locations, and people, as well as an eighth domain that was developed to include terms specific to the topic at hand—vegetable gardening.[d] To measure the length of the advice content, we used LIWC 2022 (37) to compute the number of words contained in the text (38).

To capture how unique the advice was to the writer, we undertook two text analysis approaches. First, we employed Latent Dirichlet Allocation topic modeling to analyze the co-occurrence of words within each condition. This approach allowed us to separately identify the primary topics that characterized advice written by participants in the Google condition and those in ChatGPT condition. The resulting topic distributions were then reduced in dimensionality using principal component analysis to ensure that there were the same number of topics across conditions. Using these reduced topic vectors, we calculated the average pairwise cosine similarity between all pairs of advice within each condition. The average cosine similarity measure reflects the degree of topical alignment among the advice texts, with lower scores indicating greater uniqueness of the advice in a given condition (i.e. greater divergence among the texts). To provide a convergent measure of the originality of the advice within each condition, we also calculated the average pairwise Levenshtein edit distance between all pairs of advice within each condition, with higher scores pointing to greater semantic uniqueness or greater dissimilarity in the words used (39). We predicted that advice written after learning from an LLM (vs. web search) would contain fewer words, fewer references to facts, and would be less original or unique to the writer.

Results

First, the results support the predicted effects of the search platform on the amount of effort invested in learning from search, as well as on the depth of knowledge participants reported acquiring from that search. As expected, participants who used ChatGPT spent less time on the search task than those using Google search [seconds: $M_{Google} = 742.81$, $M_{GPT} = 585.41$; $F(1, 1{,}102) = 44.61$, $P < 0.001$], suggesting that learning from LLM syntheses involved less effort than learning from standard web search results. It is worth noting that participants in the ChatGPT condition submitted a similar number of queries on average as those in the Google condition—with 2.06 prompts submitted to ChatGPT and 2.15 queries submitted to Google on average [$F(1, 870) = 0.71$, $P = 0.401$]—implying that the lower amount of time participants spent during the ChatGPT (vs. Google) search task was not driven by lower interactivity with ChatGPT compared with Google, but rather by less engagement with the search results.

Importantly, LLM use also suppressed participants' reported depth of learning on the topic compared with traditional web search: those who used ChatGPT (vs. Google) to learn about the topic at hand (how to plant a vegetable garden) reported that they learned fewer new things about the subject [$M_{Google} = 3.86$, $M_{GPT} = 3.43$; $F(1, 1{,}102) = 36.04$, $P < 0.001$], felt a lower sense of personal ownership over the knowledge they gained from their search [$M_{Google} = 3.55$, $M_{GPT} = 3.36$; $F(1, 1{,}102) = 6.82$, $P = 0.009$], and thought their search yielded less comprehensive information about the topic [$M_{Google} = 4.27$, $M_{GPT} = 4.02$; $F(1, 1{,}102) = 20.84$, $P < 0.001$].

The results also offer evidence for the predicted effects on how invested participants were in forming their advice, as well as on the content of that advice. First, as expected, those who learned through ChatGPT (vs. Google) subsequently reported putting less thought and effort into creating their advice [$M_{Google} = 4.13$, $M_{GPT} = 4.00$; $F(1, 1{,}102) = 4.94$, $P = 0.026$], reported feeling a slightly lower sense of personal ownership over the advice they created

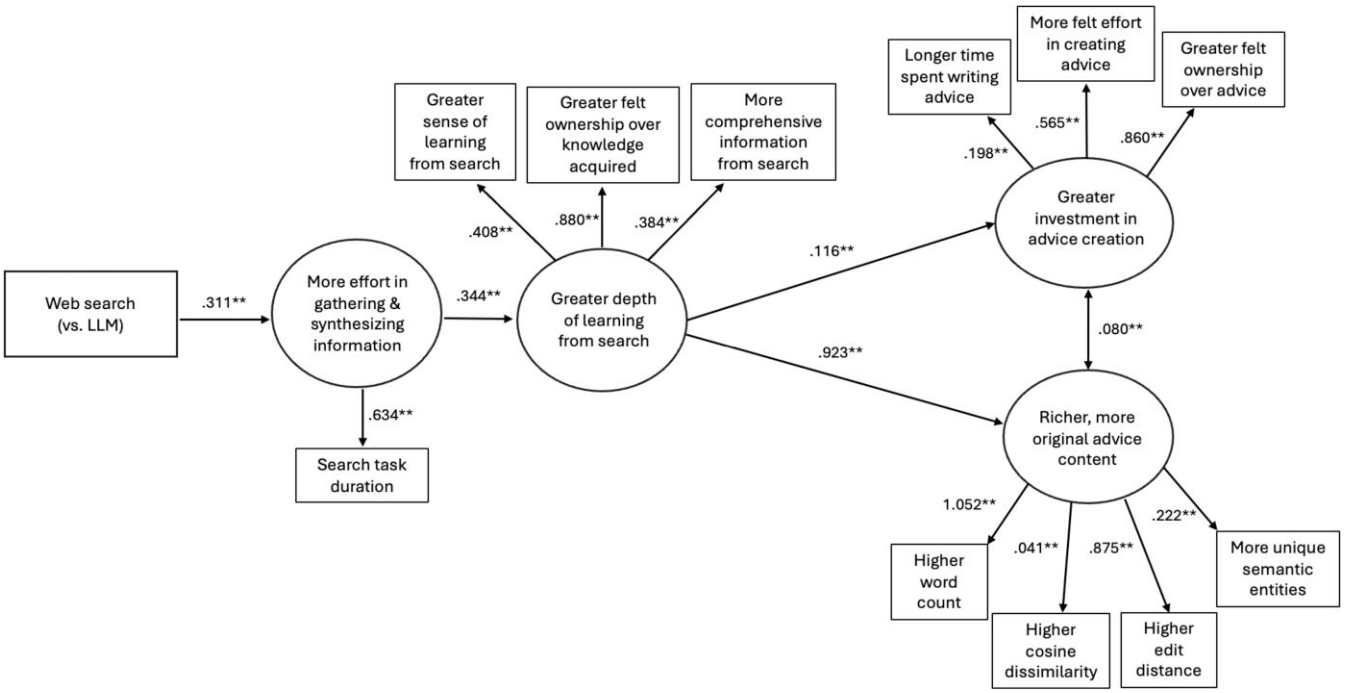

Fig. 3. Structural equation model of the theorized process (with parameter estimates) of experiment 1. **$P < 0.01$.

[$M_{Google} = 3.69$, $M_{GPT} = 3.56$; $F(1, 1{,}102) = 3.78$, $P = 0.052$], and spent less time on average writing their advice [seconds: $M_{Google} = 246.81$, $M_{GPT} = 221.40$; $F(1, 1{,}102) = 4.91$, $P = 0.027$].

Second, the content of advice written after searching through ChatGPT (vs. Google) contained linguistic markers suggestive of shallower learning. As predicted, advice written after learning from ChatGPT (vs. Google) was sparser, containing fewer words [$M_{Google} = 94.64$ words, $M_{GPT} = 84.58$; $F(1, 1{,}102) = 7.42$, $P = 0.007$] and fewer references to facts [unique entities: $M_{Google} = 0.718$, $M_{GPT} = 0.464$; $F(1, 1{,}102) = 6.66$, $P = 0.010$]. Likewise, the content of the advice written by those who learned from ChatGPT (vs. Google) was less unique to the writer, with the pieces of advice from the ChatGPT condition displaying greater topical similarity [cosine similarity: $M_{Google} = 0.057$, $M_{GPT} = 0.159$; $F(1, 1{,}102) = 168.46$, $P < 0.001$] and lower semantic uniqueness [Levenshtein edit distance: $M_{Google} = 531.86$, $M_{GPT} = 481.59$; $F(1, 1{,}102) = 13.67$, $P < 0.001$].

Finally, to test the proposed process by which ChatGPT (vs. Google) ultimately affects the advice formulated based on one's search (Fig. 1), we estimated a structural equation model using SAS's Proc Calis (40). The model provided a good fit to the data (Bentler's CFI: 0.903; standardized RMR: 0.079) and supported the theorized mechanism. As illustrated in Fig. 3, compared with those who used ChatGPT, participants who used Google search invested more effort in gathering and synthesizing information, as manifested in more time spent on the search task. This greater search effort led participants to develop deeper knowledge from their search, as manifested in greater felt learning from the search results, greater felt ownership over the knowledge they acquired, and thinking that the search results contained more comprehensive information on the subject. In turn, this greater reported depth of knowledge from web (vs. LLM) search both increased how invested participants were in creating their advice—as manifested in dedicating more thought and effort to its creation, feeling more personal ownership over their advice, and spending more time writing it—and enhanced the richness and uniqueness of the advice content—as manifested in higher word count, more references to facts, higher topical dissimilarity (cosine dissimilarity), and greater semantic uniqueness (Levenshtein edit distance).

Discussion

Experiment 1 offered initial support for the hypothesis that the less effortful and involved process of learning through LLM summaries (vs. web links) can lead to shallower knowledge about a topic, which was subsequently reflected both in their experience when writing their advice and in the content of that advice.

We also note that we conducted two preregistered replications of experiment 1 (reported in the Supplementary Material due to space limitations). In the first ($n = 2{,}402$), we asked participants to learn about a topic that they confirmed was of high personal relevance to them—how to lead a healthier lifestyle—and randomly assigned them to either the same LLM or web search conditions as in experiment 1, or to a new, third condition where the LLM syntheses were augmented by real-time web links. As expected, the results were obtained even when participants were naturally motivated to learn about the topic and even when the LLM syntheses included real-time web links—which is perhaps unsurprising given that only 26% of those participants chose to click on those links. In the second ($n = 1{,}976$), we confirm the robustness of the effects observed in experiment 1 to another consequential real-world search topic: what to do if one is the victim of a financial scam (e.g. stolen credit card information, investment fraud).

Finally, a natural concern with these findings is that because participants conducted actual Google and ChatGPT searches, the observed effects may have accrued from systematic differences in the information contained in the platforms' search results. To address this, in the second preregistered experiment, we employed the same basic procedure as in experiment 1, but rather than conducting real online searches, participants engaged in a simulated Google or ChatGPT search. This design element afforded us greater

experimental control, enabling us to directly manipulate the format in which the search results were presented while holding the information in those results constant across conditions.

Experiment 2: holding information in the search results constant

Methods

Participants were 2,016 members[e] of the Prolific panel who were randomly assigned to one of two between-subjects conditions (Google vs. ChatGPT) in a preregistered experiment (M_{age} = 41.28, SD = 17.64; 49% female, 50% male, 1% nonbinary). We excluded 37 participants for failing an attention check, yielding a final sample of 1,979. The preregistration for the study is available at https://aspredicted.org/3rxx-vktc.pdf.

As in experiment 1, participants were asked to imagine that a friend was seeking advice on how to plant a vegetable garden, and that before forming their advice, they wanted to conduct an online search on the subject using Google or ChatGPT (depending on the condition). However, whereas participants in the first experiment conducted actual searches, in experiment 2 participants were asked to type just one query to learn more about the topic either into a simulated ChatGPT interface or a simulated Google search bar. After submitting their query, they were taken to a new page where they were shown simulated results formatted either as a summary of suggestions (ChatGPT condition) or linked websites (Google condition).

Critically, while the results format differed, the information contained in the search results was identical across conditions. To achieve this, we first submitted a prompt to GPT-4o asking it to provide advice on how to plant a vegetable garden, which yielded a 291-word response consisting of seven suggestions, such as: "Choose the Right Location: Your vegetable garden should be in a spot that receives at least six hours of sunlight per day. Avoid areas that are overly windy or have poor drainage." This 291-word synthesis was used as the simulated response in the ChatGPT condition (see Supplementary Material for the full survey stimulus).

To create equivalent results for the Google condition, we submitted the same 291-word summary back to GPT-4o and asked it to create six different article variations that all retained the same set of facts as in the seven suggestions, but with each one written in the style of an article from a different media outlet, such as *Better Homes and Gardens*, *Martha Stewart*, and *The New York Times* (see Supplementary Material for the instructions submitted to GPT-4o and resulting articles). We then used these articles for the simulated results in the Google condition, which took the form of six links to webpages displaying one of the articles (see Supplementary Material for survey instrument). This experimental design achieved two goals. First, it ensured that, regardless of whether participants in the Google condition clicked just one or all six of the links, they would be exposed to the same seven suggestions on how to plant a vegetable garden as participants in the ChatGPT condition, just written in different styles. Second, by holding the set of core facts presented to participants constant across conditions, we were able to isolate the differentiating features of standard web search (vs. LLMs) that are of greatest theoretical interest: the need to actively engage with information sources by navigating web links and interpreting and distilling the information contained in the links for oneself.

After completing the search task, as in experiment 1, participants were instructed to proceed to a separate page and write their advice to a friend, and then they responded to the same set of items as in experiment 1. Timers were embedded throughout the survey to capture the amount of time spent typing the query, the amount of time spent engaging with the search results (the proxy for learning effort), and the amount of time spent writing their advice (a proxy for investment in advice formation). Finally, once data collection was completed, we subjected the advice written by participants to the same battery of NLP tools as in experiment 1 to measure the linguistic markers of depth of knowledge.

Results

The results of experiment 2 lend convergent support for our predictions, conceptually replicating experiment 1. Participants who learned through the simulated ChatGPT (vs. Google) appeared to exert less effort in learning: although they spent more time formulating their query [seconds: M_{Google} = 49.42, M_{GPT} = 59.45; $F(1, 1{,}977) = 16.35$, $P < 0.001$], they spent considerably less time engaging with the results yielded by that query [seconds: M_{Google} = 124.32, M_{GPT} = 83.65; $F(1, 1{,}977) = 58.75$, $P < 0.001$]. Notably—even though the information in the search results was held constant across conditions—participants in the ChatGPT condition still felt they learned less than those in the Google condition. Specifically, participants who learned about how to plant a vegetable garden from the simulated ChatGPT (vs. Google) results reported learning fewer new things on the subject [M_{Google} = 3.96, M_{GPT} = 3.71; $F(1, 1{,}977) = 20.70$, $P < 0.001$] and feeling a lower sense of ownership over the knowledge they gained [M_{Google} = 3.66, M_{GPT} = 3.41; $F(1, 1{,}977) = 22.66$, $P < 0.001$], though here participants felt the comprehensiveness of the information was similar across the two conditions [M_{Google} = 4.25, M_{GPT} = 4.30; $F(1, 1{,}977) = 1.25$, $P = 0.264$].

As predicted, participants were also less invested in forming their advice on how to plant a vegetable garden after learning about it from ChatGPT (vs. Google), and created advice that was suggestive of shallower knowledge on the subject. First, participants in the ChatGPT (vs. Google) condition reported exerting less thought and effort in creating their advice [M_{Google} = 4.11, M_{GPT} = 3.85; $F(1, 1{,}977) = 33.74$, $P < 0.001$], reported feeling a lower sense of personal ownership over their advice [M_{Google} = 3.75, M_{GPT} = 3.64; $F(1, 1{,}977) = 5.12$, $P = 0.024$], and spent less time writing it on average [seconds: M_{Google} = 208.84, M_{GPT} = 185.19; $F(1, 1{,}977) = 8.81$, $P = 0.003$]. Second, the results of the text analyses confirmed that advice written by participants in the ChatGPT (vs. Google) condition was sparser—containing fewer words [M_{Google} = 74.22 words, M_{GPT} = 64.49; $F(1, 1{,}977) = 22.78$, $P < 0.001$] and references to facts [M_{Google} = 4.61, M_{GPT} = 4.00; $F(1, 1{,}977) = 29.05$, $P < 0.001$]—and was less original or unique to the writer, as reflected in the advice texts having higher topical similarity [cosine similarity: M_{Google} = 0.072, M_{GPT} = 0.224; $F(1, 1{,}977) = 330.76$, $P < 0.001$] and lower semantic uniqueness within the ChatGPT condition [Levenshtein edit distance: M_{Google} = 423.98, M_{GPT} = 346.58; $F(1, 1{,}977) = 105.66$, $P < 0.001$]. Finally, the results of the same mediation model estimated in experiment 1 (reported in the Supplementary Material due to space constraints) lend convergent evidence for the theorized process (Fig. 1).

Discussion

Experiment 2 offered a conservative test of our predictions, confirming that the effects arose even when the set of facts provided by the two platforms was held constant. In particular, merely presenting that same information as an LLM summary rather than a

series of web links led participants not only to exert less effort in learning but also to report developing shallower knowledge on the topic.

While we find evidence for our thesis that the greater depth of learning from Google (vs. ChatGPT) arose due to the more effortful process it entailed—deciding which websites to click on, processing and interpreting the content of those sites, and distilling what they learned on their own—another design feature that may have contributed to this effect was that participants were able to see the same information across multiple websites (25, 26). While repeated exposure to the information likely also played a role in the deeper learning observed in the Google condition, an ex post analysis (reported in the Supplementary Material due to space constraints) revealed that the predicted effects were sustained after controlling for the number of websites visited.

Finally, it is possible that Google and ChatGPT differ in ways other than just the format of their search results that could have contributed to the observed differences. For example, most participants would have had extensive experience with using and trusting Google search as their predominant (or sole) tool for gathering information online, whereas they naturally would have much less experience with newer platforms like ChatGPT. As such, it is possible that the relative novelty of the ChatGPT (vs. Google) interface could have impeded learning. To control for such potential differences, in experiment 3, we leveraged the "AI Overview" extension in Google, which presents at the top of the standard Google search results page an LLM synthesis of the results. This feature enabled us to hold the search platform constant while varying whether participants learned from standard web results versus an LLM synthesis. Hence, experiment 3 provided a robustness test of the effects to a different LLM, this time in a laboratory setting.

Experiment 3: holding the search platform constant

Methods

Two hundred and fifty-three members[f] of the behavioral lab panel at an east-coast university were recruited to complete the experiment in a laboratory setting. Three participants were excluded for failing an attention check, leaving a final sample of 250 (M_{age} = 22.75, SD = 8.65; 73% female, 27% male). Participants were randomly assigned to one of two between-subjects conditions (Google standard vs. Google LLM). Those assigned to the Google-LLM condition were seated at a computer where the "AI Overview" extension had been installed in the Google Chrome browser, while those assigned to the Google-standard condition were seated at a computer where the extension was not installed in Chrome. For all participants, the Qualtrics survey was preopened in Google Chrome when participants sat down, such that they could begin the study when they were ready.

Similar to the prior experiments, participants were first told to imagine that they were asked by a friend for advice on a particular subject—here, how to lead a healthier lifestyle—and that they wanted to gather more information on the topic before writing their advice (see the Supplementary Material for the survey instrument). As in experiment 1, participants were instructed to conduct actual searches, but here they did so by opening a separate tab in Chrome that automatically opened to the Google search bar. Participants in the Google-LLM condition were instructed that, after submitting their query, they should focus only on the "AI Overview" summary presented at the top of the search results page without browsing through the search links presented beneath it, while those in the web search condition were asked to browse through the links and were not shown the AI overview. Similar to experiment 1, participants could interactively engage with their assigned search tool and were free to enter as many queries as they wished. Since we could not directly observe their behaviors on the Google website, a timer was embedded in the search-task instructions page of the survey—which participants kept open in a separate tab while searching on Google—which allowed us to measure the amount of time they spent searching (the proxy for learning effort).

When they felt that they had learned enough from their search to form their advice, participants were instructed to return to the survey tab and proceed to write advice for their friend. As in the prior experiments, an embedded timer captured how long participants spent writing their advice, which served as a proxy measure for how invested they were in forming their advice. After writing the advice, they responded to the same items as in the prior experiments, and once data collection was completed, participants' advice content was submitted to the same battery of human and automated text analyses.

Results

The results confirm the robustness of the basic effects of holding the search platform constant.[g] As in the prior experiments, participants appeared to exert less effort in acquiring information from Google's LLM (vs. web search), based on the amount of time they spent on the search task [seconds: $M_{standard\text{-}Google}$ = 67.37, $M_{LLM\text{-}Google}$ = 55.50; $F(1, 248) = 7.32$, $P = 0.007$]. Participants in the Google LLM (vs. web search) condition also reported that they learned less about the topic of how to lead a healthier lifestyle, reporting that they were exposed to less comprehensive information on the subject [$M_{standard\text{-}Google}$ = 4.18, $M_{LLM\text{-}Google}$ = 3.32; $F(1, 248) = 46.29$, $P < 0.001$] and learned slightly fewer new things about it [$M_{standard\text{-}Google}$ = 2.68, $M_{LLM\text{-}Google}$ = 2.41; $F(1, 248) = 2.98$, $P = 0.086$], though here they showed no difference in how much ownership they felt over the knowledge they acquired [$M_{standard\text{-}Google}$ = 3.18, $M_{LLM\text{-}Google}$ = 3.00; $F(1, 248) = 1.66$, $P = 0.199$].

As in the prior experiments, we also found evidence that after learning about a topic via an LLM (vs. web search), participants felt less invested in forming advice about it, and wrote advice that was indicative of shallower knowledge on the subject. Specifically, those who learned using Google's LLM (vs. web search) reported putting less thought and effort into writing their advice [$M_{standard\text{-}Google}$ = 3.60, $M_{LLM\text{-}Google}$ = 3.20; $F(1, 248) = 9.12$, $P = 0.003$], felt less ownership over the advice they had written [$M_{standard\text{-}Google}$ = 3.47, $M_{LLM\text{-}Google}$ = 3.06; $F(1, 248) = 8.95$ $P = 0.003$], and spent less time writing their advice on average [seconds: $M_{standard\text{-}Google}$ = 159.08, $M_{LLM\text{-}Google}$ = 134.99; $F(1, 248) = 5.55$, $P = 0.019$].

Likewise, the results of the text analyses confirmed that advice written by those who learned from Google's LLM (vs. web search) was shorter [$M_{standard\text{-}Google}$ = 90.17, $M_{LLM\text{-}Google}$ = 65.17; $F(1, 248) = 20.14$, $P < 0.001$], referenced slightly fewer facts [$M_{standard\text{-}Google}$ = 0.624, $M_{LLM\text{-}Google}$ = 0.376; $F(1, 248) = 2.80$, $P = 0.096$], and was less unique to the writer, as indicated by greater topical similarity [cosine similarity: $M_{standard\text{-}Google}$ = 0.073, $M_{LLM\text{-}Google}$ = 0.214; $F(1, 248) = 52.50$, $P < 0.001$] and lower semantic dissimilarity among the pieces of advice in the LLM condition [Levenshtein edit distance: $M_{standard\text{-}Google}$ = 502.52, $M_{LLM\text{-}Google}$ = 368.64; $F(1, 248) = 56.78$, $P < 0.001$].

Discussion

The results of experiment 3 confirm that—even when holding constant the search platform in a laboratory setting—participants who learned through LLM syntheses (vs. web links) reported developing shallower knowledge on the topic at hand, resulting in demonstrable differences in the content of their advice. That being said, the results so far have left unanswered a critical question: Do these differences in advice content yield meaningful effects on how informative and persuasive recipients find the advice to be? In the final study, we test for the downstream impacts on advice adoption.

Experiment 4: downstream effects on advice adoption

Method

Participants were 1,501 members[h] of the MTurk Connect panel who were asked to read and rate two randomly selected advice texts written by participants in experiment 3 on the topic of how to lead a healthier lifestyle. Specifically, participants were first asked to imagine that they reached out to two friends for advice on how to live a healthier lifestyle and that each friend subsequently provided advice based on research they had conducted. After reading this cover story, participants were presented with the two pieces of advice from experiment 3 in randomized order: one was randomly selected from the 125 texts written by those who learned through standard Google web links; the other from the 125 texts written by those who learned through Google's AI Overview summary. Participants were blind to the condition in which the advice was written and were asked to rate each one along a set of dimensions on a scale of 1 = "Strongly disagree" to 5 = "Strongly agree": "I find this advice very helpful," "This advice is very informative about the topic," "It seems my friend put a lot of effort into formulating this advice," "I trust this advice," "I would be very likely to adopt this advice," and "I would be very likely to recommend this advice to others." After rating each piece of advice separately, participants were then shown the two pieces of advice side-by-side and were asked to rate which one they found to be more helpful on a scale of 1 = "The advice from my first friend was more helpful," 2 = "I found both of their advice similarly helpful," and 3 = "The advice from my second friend was more helpful."

Finally, to confirm that receivers found the topic of the advice (how to lead a healthier lifestyle) to be of personal relevance, participants rated three items on a scale of 1 = "Strongly disagree" to 5 = "Strongly agree"—"How personally relevant to you is the issue of leading a healthier lifestyle?"; "Prior to this study, how interested have you recently been in learning about how to lead a healthier lifestyle?"; and "In general, to what extent do you feel you'd benefit from learning about how to lead a healthier lifestyle?"—which were averaged into a "personal relevance" index ($\alpha = 0.806$). Eight participants were excluded for failing two attention checks, resulting in a final sample of 1,493 participants ($M_{age} = 39.79$, SD = 12.85; 52% female, 46% male, 2% nonbinary). The preregistration for the study is available at https://aspredicted.org/6py8-ggdn.pdf.

Results

A preliminary analysis confirmed that most receivers viewed the topic of the advice they evaluated (how to lead a healthier lifestyle) to be highly personally relevant. For example, the mean rating across participants on the personal relevance index was 4.00 on the 5-point scale (SD = 0.801), and when posed with the item, "In general, to what extent do you feel you'd benefit from learning about how to lead a healthier lifestyle?," 77% of participants gave ratings of 4 or 5.

Next, the evaluations of the advice were subjected to a series of repeated-measures ANOVAs. The results confirmed that receivers—who were blind to the originating platform—found advice written by participants who learned from Google's AI Overview (vs. web links) to be less helpful [$M_{standard\text{-}Google} = 3.816$, $M_{LLM\text{-}Google} = 3.553$; $F(1, 1{,}492) = 85.013$, $P < 0.001$] and less informative [$M_{standard\text{-}Google} = 3.575$, $M_{LLM\text{-}Google} = 3.224$; $F(1, 1{,}492) = 103.863$, $P < 0.001$], and they believed less effort had been put into writing the advice [$M_{standard\text{-}Google} = 3.421$, $M_{LLM\text{-}Google} = 3.024$; $F(1, 1{,}492) = 94.471$, $P < 0.001$]. Likewise, receivers found advice written after learning from AI Overview (vs. web links) to be less trustworthy [$M_{standard\text{-}Google} = 4.171$, $M_{LLM\text{-}Google} = 3.906$; $F(1, 1{,}492) = 126.078$, $P < 0.001$], were less likely to recommend it to others [$M_{standard\text{-}Google} = 3.821$, $M_{LLM\text{-}Google} = 3.481$; $F(1, 1{,}492) = 117.375$, $P < 0.001$], and were less willing to adopt it themselves [$M_{standard\text{-}Google} = 3.897$, $M_{LLM\text{-}Google} = 3.711$; $F(1, 1{,}492) = 48.481$, $P < 0.001$]. Consistent with this, when comparing the advice texts directly to each other, receivers still felt that the advice written after learning from the LLM summary (vs. web links) was less helpful: a χ^2 test of choice proportions rejected a null hypothesis of equality [$\chi^2(2) = 201.156$, $P < 0.001$], with the observed proportion of LLM-based advice [50.6%, $n = 756$; 95% CI = (0.481, 0.532)] being significantly lower than that of advice based on web links [24.6%, $n = 367$; 95% CI = (0.226, 0.271)]. Finally, in the Supplementary Material, we report the results of a preregistered conceptual replication of experiment 4 ($n = 1{,}258$) that confirm the generalizability of these findings.

Taken together, these results suggest that the textual differences in the advice written by those who learned from LLM summaries (vs. web search links) observed in experiments 1–3—such as differences in linguistic richness and originality—can yield important downstream consequences for how that advice will be received by others.

General discussion

One of the most significant changes in online search in recent years has been the widespread adoption of LLMs that summarize vast arrays of information in response to a user prompt. While searching through LLMs can undoubtedly make it easier to acquire information, might this be at the detriment of learning compared with traditional web search? This work argues that a fundamental difference between LLMs and web search—the presentation of results as syntheses of information rather than web links to sources—affects the amount of effort one puts into constructing knowledge about a given topic and, thus, how deeply they learn about it from their search. We propose that compared with gathering and synthesizing information via standard search links, the lower effort involved in gathering information from LLM syntheses can lead people to develop shallower knowledge on a topic in certain contexts. When they subsequently form advice on the topic based on what they learned from their LLM (vs. web) search, people are less personally invested in forming their advice, and write advice that is shorter, contains fewer references to facts, is less original to them, and is ultimately less likely to be adopted by recipients.

We tested these predictions in an experimental context wherein participants were randomly assigned to learn about a given topic either from traditional web links or LLM syntheses and were then

asked to provide advice for others based on what they learned. Results from seven online and lab experiments—including four reported in the main text and three conceptual replications in the Supplementary Material—lent convergent support for our predictions. The effects were robust to participants conducting *actual* LLM and web searches (experiments 1 and 3), simulated searches where the content of the results was controlled (experiment 2), and the use of different LLMs (ChatGPT, Google's AI Overview). The effects were also robust to a range of search topics, from how to plant a vegetable garden (experiments 1 and 2), to how to lead a healthier lifestyle (experiment 3, Supplementary Material), to what one should do as the victim of a financial scam (Supplementary Material). Importantly, we found that the observed differences were not merely driven by differences in the information available from ChatGPT versus Google. Participants reported developing shallower knowledge from LLM summaries versus web search links even when the underlying engine (Google) was held constant (experiment 3), even when the facts in the results were held constant and only the presentation format of the results varied (experiment 2), and even when the LLM syntheses were augmented by real-time web links (Supplementary Material).

Taken together, our findings suggest that one of the major benefits of LLM syntheses—sparing users the need to browse through results and synthesize information themselves—can come at the expense of developing deeper knowledge on a subject in certain contexts. In this sense, one might view learning through LLMs rather than traditional web search as, at times, analogous to being shown the solution to a math problem rather than trying to solve it oneself.

Implications, limitations, and future research

Recent years have witnessed a growing philosophical debate over the risks of AI use, such as the possibility that increased reliance on AI will diminish humans' ability to complete tasks themselves (41–43). While this has always been a possible byproduct of technological advancement (44), some argue that the deskilling that could arise from AI use will be especially pernicious by degrading basic thinking and reasoning skills over time (43, 45). Consistent with this, recent findings have shown, for example, that the use of AI as an aid can constrain the range of creative solutions to specific problems (46). Nevertheless, whether AI use can have such far-reaching effects remains an open empirical question.

Our work contributes to this debate by offering experimental evidence of the effects of AI in the context of learning and advice formation. We find that unlike learning from web search, which requires one to comprehend, interpret, and integrate information for oneself, learning from presynthesized LLM results can lead to the development of shallower knowledge on a subject. Hence, it is possible that increased reliance on LLM syntheses for learning in lieu of standard web search, especially among younger people, threatens to "deskill" the ability to engage in active learning—a critical ingredient in knowledge development (9, 33).

That said, our findings naturally come with important caveats. For one, although participants in most of our experiments interacted with *actual* LLMs (e.g. GPT-4o) and web search tools (Google), the results arose in the context of a specific experimental task: that of learning about a topic with the intention to share advice on it with others. While we argue that this is a common use case for LLMs, different applications of LLMs may naturally yield different results. For example, there would presumably be no obvious advantage to using traditional web search (vs. LLMs) if one's goal is simply to look up a particular factual answer (such as a historical date). Nonetheless, while LLMs might in general be a more efficient way to acquire *declarative* knowledge, or knowledge on specific facts (47), our findings suggest that they may not be the best tool for developing deep *procedural* knowledge about a topic—that is, an understanding of how things work, and an ability to create new knowledge.

Further, our data came from task contexts where deep prior knowledge was not needed to comprehend basic information on the topic (e.g. how to live a healthier lifestyle, how to plant a vegetable garden). The effects might not generalize, however, to learning in highly specialized domains—such as advanced science or math topics—where LLM syntheses might better help users comprehend and interpret complex topics compared with web links. Likewise, it is important to note that our evidence that users develop shallower knowledge from LLM summaries (vs. web links) centered on participants' self-reported lower depth of learning, the shorter and less factual advice text they formed, as well as recipients' perceptions that their advice was less informative. That said, there may well be other metrics of learning—such as performance on standardized tests—where LLMs offer advantages over web search, for instance, by better guiding users to the key facts essential for test preparation. An important area for future work will be to more thoroughly explore the boundaries and generalizability of the observed differences to other learning contexts and metrics.

Another avenue for further research would be to systematically explore which specific LLM models and features are more conducive to learning. For example, future work could examine whether the effects observed here generalize to deep research tools available through LLMs like Perplexity.ai, Gemini, and ChatGPT, which perform in-depth research and analysis on behalf of users. While use of these advanced tools may at times enhance knowledge development compared with synthesizing original sources for oneself via web links, our findings suggest that outsourcing research to such tools may amplify, rather than attenuate, the differences observed here, at least in certain contexts.

In sum, we see our findings as carrying a message of caution for the growing reliance on LLMs for knowledge acquisition. One of the risks of relying on LLM summaries in lieu of traditional web search links is that it can transform learning from a more active to passive activity—which has been shown to yield inferior learning outcomes in other settings (48, 49). We thus believe that while LLMs can have substantial benefits as an aid for training and education in many contexts, users must be aware of the risks—which may often go unnoticed—of overreliance. Hence, one may be better off not letting ChatGPT, Google, or another LLM "do the Googling."

Notes

[a] This sample size is based on a G*Power analysis that would provide a 90% chance of detecting a small effect ($d = 0.10$) at $P < 0.05$.

[b] For the Google condition, we used ngrok (https://ngrok.com/) to enable the Google site running and Google's Custom Search JSON API (https://developers.google.com/custom-search/v1/overview) to handle search queries.

[c] For the ChatGPT condition, we used OpenAI's Chat Completion API (https://platform.openai.com/docs/api-reference/chat) and the GPT-3.5-turbo model to generate responses to user prompts.

[d] The added domain included terms such as "soil," "water," "fertilizer," "compost," "carrots," "tomatoes," "seeds," "plants," "mulch," "vegetables," "tools," "garden," and "gardening."

[e] The sample size was again determined based on a G*Power analysis that would provide a 95% chance of detecting a small effect ($d = 0.15$) at $P < 0.05$.

[f]The sample size was the largest feasible under lab facility constraints and subject pool.

[g]In experiment 3, we could not run the same mediation model as in experiments 1 and 2 because the maximum likelihood estimates of the SEM were not estimable due to the limited sample size of this laboratory experiment.

[h]This sample size is based on a G*Power analysis that would provide an 80% chance of detecting a small effect ($d = 0.15$) at $P < 0.05$.

Acknowledgments

The authors thank Shawn Zamechek for programming the in-house ChatGPT in experiment 1, and Hal Hershfield for collaborating on an early version of the project.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work was funded with support from the Wharton Dean's Office and the Wharton Behavioral Lab.

Author Contributions

Shiri Melumad (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing) and Jin Ho Yun (Data curation, Formal analysis, Methodology, Investigation, Resources, Software, Validation, Visualization, Writing—review & editing)

Preprints

The original version of the manuscript was posted on a preprint at https://ssrn.com/abstract=5104064.

Data Availability

All relevant data, surveys, and preregistrations used in analysis can be accessed via OSF.

References

1 J. Kaddour, *et al.* Challenges and applications of large language models [preprint], 2023, arXiv:2307.10169, http://arxiv.org/abs/2307.10169 [Accessed 2025 Jan 10].
2 Meincke L, Mollick ER, Terwiesch C. Prompting diverse ideas: increasing AI idea variance [preprint], 2024, arXiv:2402.01727, http://arxiv.org/abs/2402.01727 [Accessed 2025 Jan 10].
3 Ray PP. 2023. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst*. 3:121–154.
4 Naveed H, *et al.* A comprehensive overview of large language models [preprint], 2024, arXiv:2307.06435, http://arxiv.org/abs/2307.06435 [Accessed 2025 Jan 10].
5 Scharf A. Google's AI Overviews: 4 key performance insights. 2024. https://www.seerinteractive.com/insights/google-ai-overviews-performance-insights [Accessed 2025 Jan 10].
6 Chapekis A, Lieb A. Google users are less likely to click on links when an AI summary appears in the results. *Pew Research Center*. https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/ [Accessed 2025 Sept 2].
7 Stadler M, Bannert M, Sailer M. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Comput Human Behav*. 160:108386.
8 Rieh SY, Collins-Thompson K, Hansen P, Lee H-J. 2016. Towards searching as a learning process: a review of current perspectives and future directions. *J Inf Sci*. 42:19–34.
9 Vakkari P. 2016. Searching as learning: a systematization based on literature. *J Inf Sci*. 42:7–18.
10 Kung TH, *et al.* 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2:e0000198.
11 Terwiesch C, Meincke L, Nave G. The AI ethicist: fact or fiction? [preprint], 2023, https://papers.ssrn.com/abstract=4609825 [Accessed 2024 Jul 31].
12 Dell'Acqua F, *et al.*, Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality [preprint], 2023, https://papers.ssrn.com/abstract=4573321 [Accessed 2025 Jan 10].
13 Noy S, Zhang W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 381:187–192.
14 Si C, Yang D, Hashimoto T. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers [preprint], 2024, arXiv:2409.04109, http://arxiv.org/abs/2409.04109 [Accessed 2025 Jan 10].
15 Lee BC, Chung J. 2024. An empirical investigation of the impact of ChatGPT on creativity. *Nat Hum Behav*. 8:1906–1914.
16 Lyu W, Wang Y, Chung T, Sun Y, Zhang Y. 2024. Evaluating the effectiveness of LLMs in introductory computer science education: a semester-long field study. Proceedings of the Eleventh ACM Conference on Learning@ Scale. p. 63–74.
17 Deng R, Jiang M, Yu X, Lu Y, Liu S. 2025. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput Educ*. 227:105224.
18 Mahapatra S. 2024. Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. *Smart Learn Environ*. 11:9.
19 Bastani H, *et al.* 2025. Generative AI without guardrails can harm learning: evidence from high school mathematics. *Proc Natl Acad Sci U S A*. 122:26.
20 J. Zhang, *et al.* Investigation of the effectiveness of applying ChatGPT in dialogic teaching of electronic information using electroencephalography. 2024 6th International Conference on Computer Science and Technologies in Education (CSTE). IEEE; 2024. p. 150–154.
21 Abbas M, Jam FA, Khan TI. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *Int J Educ Technol High Educ*. 21:10.
22 Fernández-Pichel M, Pichel JC, Losada DE. 2025. Evaluating search engines and large language models for answering health questions. *NPJ Digit Med*. 8:153.
23 Spatharioti SE, Rothschild DM, Goldstein DG, Hofman JM. Comparing traditional and LLM-based search for consumer choice: a randomized experiment [preprint], 2023, arXiv:2307.03744, http://arxiv.org/abs/2307.03744 [Accessed 2025 Jan 10].
24 Xu RR, Feng YK, Chen H. ChatGPT vs. Google: a comparative study of search performance and user experience [preprint], 2023, arXiv:2307.01135, https://papers.ssrn.com/abstract=4498671 [Accessed 2024 May 29].
25 Ghosh S, Rath M, Shah C. Searching as learning: exploring search behavior and learning outcomes in learning-related tasks. Proceedings of the 2018 Conference on Human Information

Interaction & Retrieval, CHIIR '18. Association for Computing Machinery; 2018. p. 22–31.
26 von Hoyer J, *et al.* 2022. The search as learning spaceship: toward a comprehensive model of psychological and technological facets of search as learning. *Front Psychol*. 13:827748.
27 Zhang P, Soergel D. 2014. Towards a comprehensive model of the cognitive process and mechanisms of individual sensemaking. *JASIST*. 65:1733–1756.
28 Zhang P, Soergel D. 2016. Process patterns and conceptual changes in knowledge representations during information seeking and sensemaking: a qualitative user study. *J Inf Sci*. 42:59–78.
29 Bjork EL, Bjork RA. Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning. In: *Psychology and the real world: essays illustrating fundamental contributions to society*. Worth Publishers, 2011. p. 56–64.
30 Alter AL. 2013. The benefits of cognitive disfluency. *Curr Dir Psychol Sci*. 22:437–442.
31 Cockburn A, Kristensson PO, Alexander J, Zhai S. Hard lessons: effort-inducing interfaces benefit spatial learning. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07. Association for Computing Machinery; 2007. p. 1571–1580.
32 Alter AL, Oppenheimer DM, Epley N, Eyre RN. 2007. Overcoming intuition: metacognitive difficulty activates analytic reasoning. *J Exp Psychol Gen*. 136:569–576.
33 Diemand-Yauman C, Oppenheimer DM, Vaughan EB. 2011. Fortune favors the bold (and the italicized): effects of disfluency on educational outcomes. *Cognition*. 118:111–115.
34 Alba JW, Hutchinson JW. 1987. Dimensions of consumer expertise. *JCR*. 13:411–454.
35 Garrett N. 2009. Computer-assisted language learning trends and issues revisited: integrating innovation. *Mod Lang J*. 93:719–740.
36 Tripathi A. Word2Vec and semantic similarity using spacy | NLP spaCY Series | Part 7. *Data Science Duniya*; 2020. https://ashutoshtripathi.com/2020/09/04/word2vec-and-semantic-similarity-using-spacy-nlp-spacy-series-part-7/ [Accessed 2025 Jan 12].
37 Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. *The development and psychometric properties of LIWC-22*. University of Texas at Austin, 10, 2022.
38 Dumas D, Organisciak P, Maio S, Doherty M. 2021. Four text-mining methods for measuring elaboration. *JCB*. 55:517–531.
39 Navarro G. 2001. A guided tour to approximate string matching. *ACM Comput Surv*. 33:31–88.
40 SAS Institute. *SAS 9.4 output delivery system: user's guide*. 3rd ed. SAS Institute, 2014.
41 Fügener A, Grahl J, Gupta A, Ketter W. 2021. Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI. *MIS Q*. 45:1527–1556.
42 Valenzuela A, *et al.* 2024. How artificial intelligence constrains the human experience. *J Assoc Consum Res*. 9:241–256.
43 Zhai C, Wibowo S, Li LD. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *SLE*. 11:28.
44 Wood S. 1987. The deskilling debate, new technology and work organization. *Acta Sociologica*. 30:3–24.
45 Habib S, Vogel T, Anli X, Thorne E. 2024. How does generative artificial intelligence impact student creativity? *JoC*. 34:100072.
46 Meincke L, Nave G, Terwiesch C. 2025. ChatGPT decreases idea diversity in brainstorming. *Nat Human Behav*. 9:1107–1109.
47 Anderson JR. *Language, memory, and thought*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1976.
48 Yannier N, *et al.* 2021. Active learning: "hands-on" meets "minds-on.". *Science*. 374:26–30.
49 Zosh JN, *et al. Learning through play: a review of the evidence*. LEGO Fonden, Billund, Denmark, 2017.