# Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance

Sofia Eleni Spatharioti
Microsoft Research
New York, New York, USA
s.spatharioti@gmail.com

David Rothschild
Microsoft Research
New York, New York, USA
David@ResearchDMR.com

Daniel G Goldstein
Microsoft Research
New York, New York, USA
dgg@microsoft.com

Jake M Hofman
Microsoft Research
NYC, New York, USA
jmh@microsoft.com

## Abstract

Recent advances in large language models (LLMs) are transforming online applications, including search tools that accommodate complex natural language queries and provide direct responses. There are, however, concerns about the veracity of LLM-generated content due to potential for LLMs to "hallucinate". In two online experiments, we examined how LLM-based search affects behavior compared to traditional search and explored ways to reduce overreliance on incorrect LLM-based output. Participants assigned to LLM-based search completed tasks more quickly, with fewer but more complex queries, and reported a more satisfying experience. While decision accuracy was comparable when the LLM was correct, users overrelied on incorrect information when the model erred. In a second experiment, a color-coded highlighting system helped users detect errors, improving decision accuracy without affecting other outcomes. These findings suggest that LLM-based search tools have promise as decision aids but also highlight the importance of effectively communicating uncertainty to mitigate overreliance.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI)**; • **Information systems** → *Web searching and information discovery*; **Users and interactive retrieval**.

## Keywords

large language models, information retrieval, search, uncertainty, hallucination, decision making

**ACM Reference Format:**
Sofia Eleni Spatharioti, David Rothschild, Daniel G Goldstein, and Jake M Hofman. 2025. Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3706598.3714082


*CHI '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714082

## 1 Introduction

Recent advances in artificial intelligence (AI), specifically in large language models (LLMs), are changing online tools used by billions of people. The search engine was one of the first applications to be transformed following the release of ChatGPT in November of 2022. By May 2023, Microsoft made its LLM-based search tool generally available and Google followed suit so that by the end of 2023 the two search engines with over 90% of global and US market share offered free LLM-based search [46].

From a user experience perspective, traditional web search and LLM-based search differ in a number of ways, each having their own advantages and disadvantages. When using traditional web search, users typically issue relatively succinct queries [17, 45] and are presented with a list of hyperlinks to and snippets from web pages containing relevant reference information. There are several benefits of this style of information retrieval. Traditional search allows rather direct access to source material through hyperlinks. In addition, traditional search enables users to see convergence or disagreement among distinct sources of information through the different references on a results page [12]. Traditional search is also explicitly optimized to return authoritative results [5] and provides additional cues about the reliability of information, for example through the domains and publishers of different results (e.g., information from the Library of Congress might be considered more trustworthy than one from an unknown domain).

There are, however, also several drawbacks to the traditional web search process. While it is convenient to have access to reference material from different sources, synthesizing information from them can be challenging and time consuming. Whereas relevant information is sometimes presented in the snippets or "instant answers" on a search result page [1, 10], users often have to click through to several different results and search within those respective pages to find pertinent information. In addition, verbose or complex queries can often lead to poor search results [3, 15, 28], and given that many real-world decision tasks are complex, this can result in users needing to break down a task into a series of simpler queries [16, 18, 36]. Lastly, it can be a technical challenge for search engines to retain context among sequences of such queries within a complex search session [13, 22, 27, 29].

LLM-based search has a different set of strengths and weaknesses. Among its strengths, LLM-based search provides a natural language interface capable of managing complex queries and delivering detailed, direct responses inferred from massive amounts of data. LLM-based search also lends itself to retaining more context from the search session than traditional search, enabling users to engage in a conversational exchange to refine and follow up on a sequence of queries. However, LLMs are known to have issues with "fabrication" or "hallucination" in which they generate plausible-sounding but factually inaccurate responses [30, 32]. This issue is particularly problematic in the context of LLM-based search, as it could foster overreliance if users assume that the information they are shown is always correct [8, 38]. Furthermore, compared with traditional search, LLM-based search offers fewer reliable cues for users to gauge the accuracy of information. Responses may not be grounded in or may not expose explicit hyperlinks to source materials, which users traditionally rely on to verify statements. Even when external links are provided, they are not displayed as prominently as in traditional web search and there can be discrepancies between the content of LLM-generated responses and the sources they cite [30].

How will the differences between traditional and LLM-based search affect people's every day decision making? On one hand, LLM-based search could offer substantial benefits, providing an easier-to-use interface that speeds up complex tasks to help people accomplish their goals more quickly or free up time for them to aquire more information. On the other hand, fabrications in LLM-generated results could mislead people, and so while they might complete tasks more quickly, they might also make sub-optimal decisions based on inaccurate information.

Here we present two randomized experiments to investigate this question. The first looks at how LLM-based search tools affect decision making compared to traditional search, and the second proposes and tests interventions to mitigate overreliance on erroneous LLM-based content. Participants in our experiments were asked to solve a series of decision tasks that involved researching and comparing different products, and were randomly assigned to do so with either an LLM-based search tool or a traditional search engine. In our first experiment, we find that participants using the LLM-based tool were able to complete their tasks more quickly, using fewer but more complex queries than those who used traditional search. Moreover, these participants reported a more satisfying experience with the LLM-based search tool. When the information presented by the LLM was reliable, participants using the tool made decisions with a comparable level of accuracy to those using traditional search, however we observed overreliance on incorrect information when the LLM erred. Our second experiment further investigated this issue by randomly assigning some users to see a simple color-coded highlighting scheme to alert them to potentially incorrect or misleading information in the LLM responses. Overall we find that this confidence-based highlighting substantially increases the rate at which users spot incorrect information, improving the accuracy of their overall decisions while leaving most other measures unaffected. Together these results suggest that LLM-based information retrieval tools have promise for increasing the productivity of people engaged in decision tasks, and highlights the opportunity of communicating uncertainty to help people know when to scrutinize or further verify LLM output.

In what follows we first review related work and then present the design and results of each experiment.

## 2 Related Work

In this work we explore how people use and make decisions with traditional vs. LLM-based search. Our research builds on prior studies in generative AI for knowledge work, as well as the extensive literature on how people interact with traditional search engines, in several key ways. First, unlike studies focused on writing or coding, we examine how LLMs impact search and question answering. Second, we focus on every day decision making, a broad category that, to our knowledge, has not yet been explored in the context of LLM-assisted productivity. Finally, we propose and test simple user interface solutions to mitigate overreliance on potentially unreliable text generated by LLMs.

### 2.1 LLMs and Productivity

LLMs have shown promise in enhancing producivity in several domains. For instance, Noy and Zhang [35] conducted an online experiment to evaluate the impact of LLM-based writing assistants on worker productivity and associated measures, finding that AI assistance improved productivity and enhanced the quality of writing in several ways. A separate field experiment had similar findings, demonstrating that LLM-based tools could be used to improve resume clarity and made candidate selection more efficient [48].

Brynjolfsson et al. [7] explored the influence of LLMs on productivity in the customer service sector by investigating the deployment of a GPT-based chat assistant. They discovered that it positively impacted productivity, especially for lower-skilled workers, and resulted in other beneficial outcomes (e.g., fewer escalations). In other work, Dell'Acqua et al. [11] executed a controlled experiment with consultants at Boston Consulting Group, where they showed increased productivity for certain types of tasks, with more gains for lower skilled workers. However, they also showed accuracy drops for tasks for which the LLM-based tools performed poorly.

In the domain of software developer productivity, Peng et al. [41] conducted a controlled experiment using an LLM-based coding tool (GitHub Copilot) to assess its impact on productivity. Developers who were assigned coding tasks and randomly provided with LLM assistance completed the tasks in less than half the time it took the control group. The study revealed that certain groups (e.g., less experienced developers) reaped more benefits, whereas earlier research suggested that the rate of acceptance of code suggestions, rather than their actual persistence in the final code, predicts developers' perceptions of productivity [52].

These studies demonstrate the widespread potential for LLM-based tools to improve productivity, however they also raise important questions about the reliability of LLM outputs, particularly in tasks where accuracy is critical.

### 2.2 Hallucinations and Overreliance

Despite the promise of LLMs for improving the speed and quality of knowledge work mentioned above, their use is not without risk, particularly when they produce erroneous or fabricated outputs [30, 32]. The problem of overreliance on AI has been a concern both before [31, 34, 34, 39, 43, 50] and after the current interest in

LLM-generated errors [21, 38, 49]. Addressing these issues requires both algorithmic solutions to identifying incorrect LLM-generated content and design interventions to communicate uncertainty to users to mitigate overreliance on such content [38, 49]. Conversely, unreliable output can have metacognitive effects on if and how users choose to use a system or tool [47].

There are many existing efforts to algorithmically infer the reliability of LLM-generated content. Kadavath et al. [19] demonstrated that LLMs can be trained to assign confidence levels to their responses. However, these confidence levels are not always calibrated, particularly in more complex contexts. Similarly, Yin et al. [51] found that while LLMs can signal uncertainty, their ability to accurately do so is still limited and context-dependent. To address this, Lin et al. [25] proposed training LLMs to generate text that reflects uncertainty as part of their output. These efforts mark important steps towards the goal of automatically identifying incorrect or unreliable information, and highlight that more work is needed to improve such methods.

In parallel researchers have investigated different ways of communicating uncertainty around LLM-based content to users to see how it affects their behavior. Early work by Vasconcelos et al. [49] in programming explored using color-coded highlighting to direct attention to problematic LLM-generated code. They found that highlighting code based on its likelihood of being edited (edit probability) was more effective than highlighting based on the model's internal confidence (generation probability). More recent work has explored using natural language to communicate uncertainty (e.g., "I'm not sure, but..."), finding that such approaches reduce overreliance, although the exact wording impacts effectiveness [21]. Other work has conducted user studies looking at more elaborate, interactive systems for alerting users to potential unreliable information, again finding that these approaches hold promise for reducing overreliance on incorrect information, including HILL [23], which explores answering questions built off of a standardized question set, and RELIC [9], which involves reviewing long-form content.

In our work, we adapt the simple uncertainty highlighting approach in [49] to LLM-based search, using color-coding to alert users to potentially misleading information in LLM-generated responses. Uniquely, we evaluate this technique at scale with hundreds of users to investigate how it affects their decision making and user experience compared to using traditional search.

## 2.3 The Evolution of Web Search

The study of how users interact with search engines has a long history, with foundational models introduced by Bates, who highlighted how users' skills co-evolve with interface design [2]. As interfaces have evolved, so too have user behaviors and expectations. Bennett et al. [4] offered a detailed categorization of search session elements—from query terms to session lengths—that inform our current understanding of user behavior, whereas Liu [26] provided a broad survey of what users aim to achieve in different types of search sessions, which remains relevant as users now encounter LLM-enhanced tools. Studies have also documented the impact of new search features, such as auto-completion and real-time suggestions, on user behavior [33].

LLMs represent the latest in a series of technological disruptions that are reshaping how users interact with search tools. Our research builds on this work, focusing specifically on how LLM-based search tools change decision making processes compared to traditional web search.

# 3 Domain and Research Questions

We focus on the domain of online product research, specifically on the task of purchasing an automobile [42]. Suppose that a user is in the market for an SUV that offers ample cargo space (to transport packages) but with a small total length (to facilitate parking). How might they go about searching for a vehicle that maximizes this ratio of cargo space to total length? Looking at data from a traditional search engine (see Appendix A), we see that people often search for one vehicle (e.g., a Jeep Wrangler) at a time and for information on one dimension (e.g., cargo space) at a time. A hypothetical user looking for a high-volume, low-length SUV might issue a series of simple queries when using a traditional search engine. A series of 1 product and 1 dimension searches could be: "jeep wrangler total cargo space" in one query and "jeep wrangler length" in another, followed by some calculations to figure out the ratio of cargo space to length.

However, with LLM-based tools, a user might instead issue just one prompt that directly addresses the decision they are looking to make. When choosing between two SUVs, a person might make a head-to-head comparison [6, 20] with a complex query like "Which vehicle has a larger cargo space to length ratio, a Jeep Wrangler or a Hyundai Santa Fe?"

We are interested in the questions of how quickly users will adapt to this new style of search, and how will they react if an LLM's response contains incorrect information. To gain insight into these questions, we designed and conducted two online experiments in which participants were randomly assigned to complete a series of consumer product research tasks using either an LLM-based search tool or a traditional search tool. We designed these experiments to focus on the following questions:

- **Research question 1 (efficiency):** How will task completion time and the number (and complexity) of queries issued differ between LLM and traditional search conditions?
- **Research question 2 (accuracy):** How will the accuracy of decisions differ between LLM and traditional search conditions?
- **Research question 3 (perceptions):** How will the user experience and perceived reliability of results differ between LLM and traditional search conditions?
- **Research question 4 (confidence and errors):** How will participants compensate for mistakes in the LLM responses with and without cues indicating the model's confidence in its output?

# 4 Experiment 1

We designed an experiment in which participants assume the role of running an urban delivery service and are looking to purchase vehicles to meet their business needs. The experiment consisted of a series of tasks. For every task, participants were provided with a pair of SUVs and were asked to choose the one that is the best option for making deliveries. To capture common criteria for choosing a vehicle for deliveries (ability to hold many packages and parking flexibility), we defined the main metric for choosing a vehicle as the cargo space to total length ratio. Cargo space in this case is defined as the maximum amount of space behind the driver's seat, with all other seats folded down, and total length is defined as the exterior length of the vehicle. Therefore, a higher cargo space to total length ratio would translate to a vehicle that is better suited to meet the delivery service business needs. This design ensures that there is both a correct answer to each task participants are given, and provides them with clear criteria to be used in making their decisions. An example of such task can be seen in Figure 1a. In this example, participants must choose whether the Kia Sportage or the Toyota RAV4 is the best product for the provided scenario.

Participants were invited to complete a series of five of the above defined tasks, where the goal for each task was to determine the best option from a randomly generated vehicle pair. We varied the type of assistance they received for their search in making their decisions in a two condition, between-subjects design. In one condition, participants were provided with "an experimental search engine" built using the Bing API. Similar to traditional search, the search engine returned a series of clickable links with descriptions based on the input query, that participants would be able to visit to get more information. In the second condition, participants had access to "an experimental AI-generated search tool", built using GPT 3.5.[1] The LLM-based tool responded to participant queries in natural language but lacked conversational capabilities. This design choice minimized differences to the traditional search condition, making it more likely that any observed effects could be attributed to the nature of the responses rather than the mode of interaction (conversational vs. non-conversational).

Figure 1b shows the difference in responses from the search tool for the two conditions. No information on the technology used behind either search tool was provided to participants, but in both conditions participants were given a short tutorial on how to use the corresponding tool and what to expect from it, as shown in Appendix B.1.

We imposed a limit of 10 searches per task and a limit of 1,000 characters per search. In addition, participants had to complete at least one search to be allowed to make a decision and thus proceed to the next task. For both conditions, participants had access to their full search history and could revisit search tool responses at any time. After completing all available tasks, participants were also asked to complete a brief survey about their experience to conclude the experiment.

The LLM-based search tool was given the following pre-prompt in order to provide a consistent experience for participants and to signal to participants that it is not a conversational tool: "You are a search engine to be used for finding facts about motor vehicles and doing math. If you are given a query about the features of a commercial car, truck, or SUV, do your best to answer it. If you are given a query that involves doing math, do your best to answer it. If you are given a query that seems like it's trying to refer to a previous conversation, respond with 'Sorry, I do not have the ability to refer to information from past questions or answers.' Otherwise respond with 'Sorry, that does not seem like a relevant query. Please try again.' Show your work." This pre-prompt was not visible to participants. To facilitate reproducibility, the tool was also configured to run at zero "temperature", so that it was deterministic: given a particular input query it always returned the same response. (Frequency penalty and presence penalty were also set to zero.)

In order to increase the chances that participants in the LLM-based tool condition would see a case on which the LLM-based tool reports erroneous information, we set the last task to be identical for all participants and to involve a vehicle with which the LLM-based tool tends to report the incorrect amount of cargo space (the 2020 Toyota 4Runner). In this particular case the LLM-based tool is prone to confuse the cargo space with the seats up for the cargo space with the seats down, thus reporting that the SUV with the largest cargo space (with seats down, as specified in the instructions) actually has the smallest cargo space. Participants in both conditions saw this item in the last position (task 5), making it the task where overreliance was more likely to occcur [34]. We refer to this task as "challenging", and refer to the other tasks as "easy". During the experiment design phase, we ran automated tests to evaluate the LLM's results across a range of reasonable prompts, to further ensure the distinction between "easy" and "challenging" tasks.

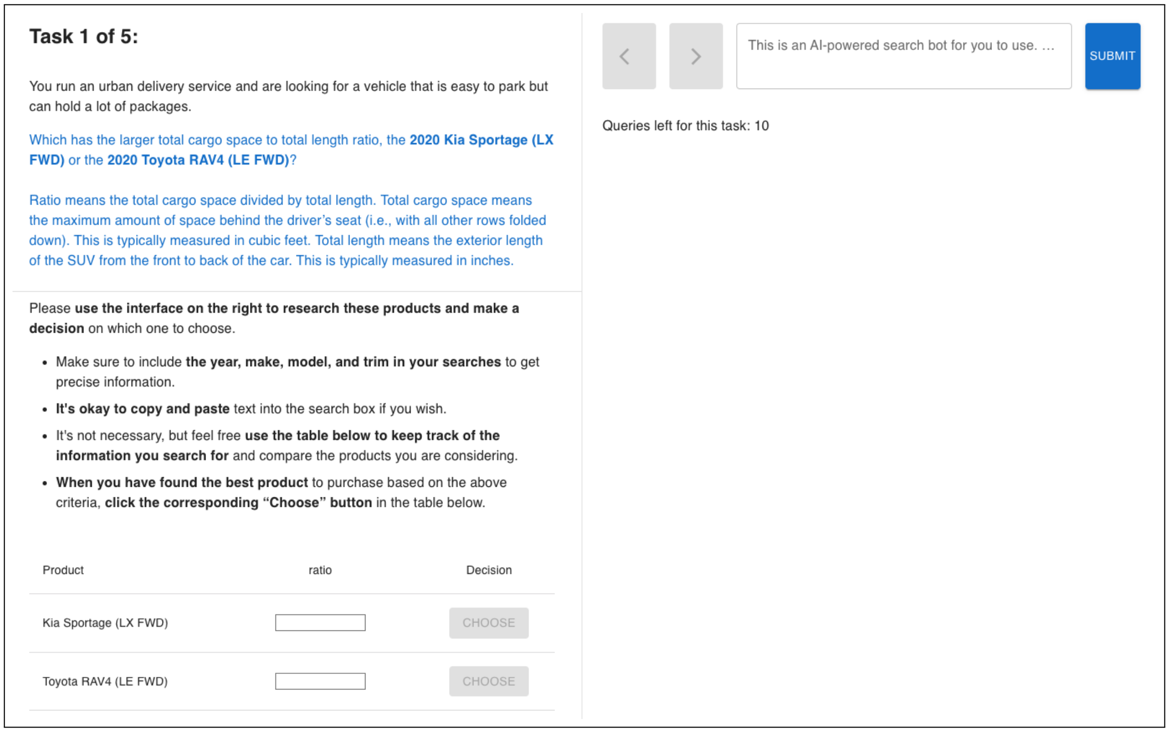
We recruited 90 U.S. based participants from Amazon Mechanical Turk. For qualifications, we required at least 2,500 HITs approved with a 99% minimum approval rate, along with an additional Masters qualification on the Mechanical Turk platform. Participants were paid \$4 for completing the experiment. We did not collect any demographic information from participants.

### 4.1 Results for Experiment 1

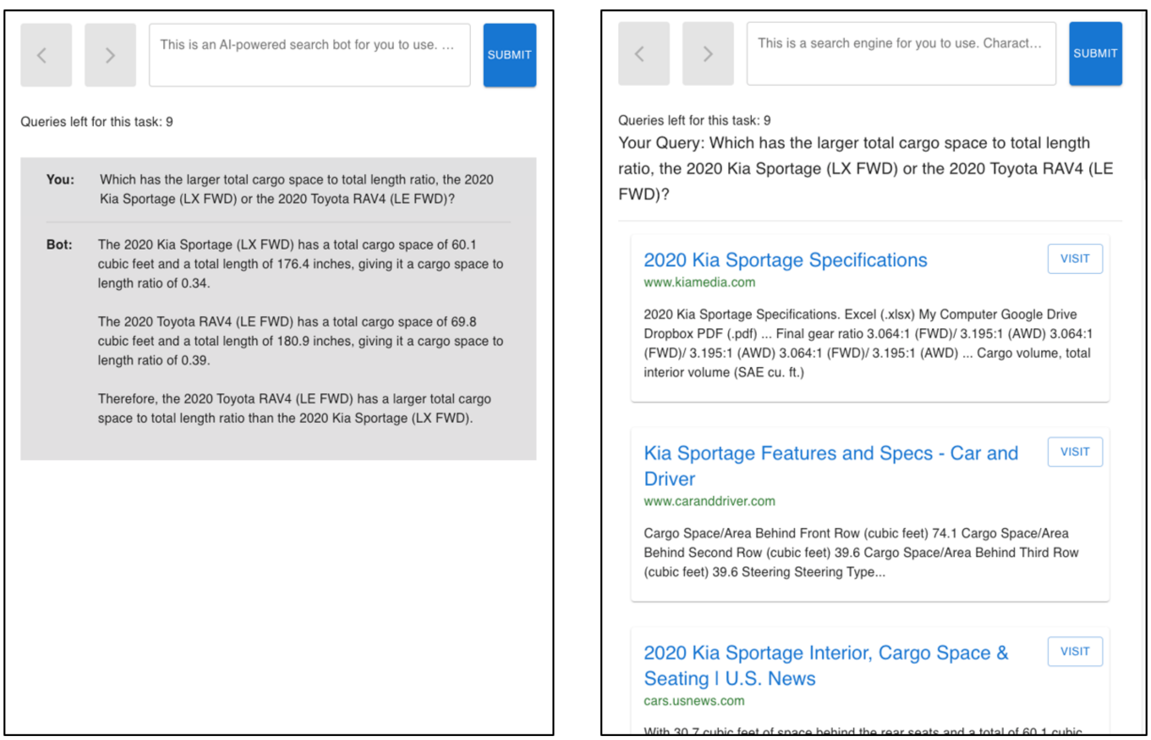
*Efficiency.* As shown in Figure 2a, participants took less time to complete the task in the LLM-based search condition relative to the traditional search condition, a pattern which is apparent as early as the first round. In both conditions we see a learning effect where participants are slower in the first task compared to subsequent tasks. Participants are simultaneously learning about the task and the domain, while also learning about the functionality of the tool they are using. In addition to the time to respond being lower on average in the LLM-based search condition, the variance was also lower.

A linear mixed model fit to task duration confirms this. Specifically, we modeled the log task time based on a random effect by participant id, controls for task number, and a fixed effect for condition (`lmer: log10(task_duration_full) ~ (1|worker_id) + as.factor(task_num) + condition`). The fixed effects estimates revealed statistically and practically significant effects of task number and condition on the log-transformed task duration. The LLM-based condition significantly reduced the log-transformed task duration compared to the traditional search condition (Estimate = -0.31613, SE = 0.05542, $t(78) = -5.70$, $p < .001$), and all tasks were faster, on average, relative to the first across conditions. The estimated average task durations, back-transformed from the log10 scale, were 3.4 minutes (95% CI [2.8, 4.1]) for the traditional search condition and 1.6 minutes (95% CI [1.4, 1.9]) for the LLM-based search condition, a roughly 50% reduction for the LLM-based tool.

[1]At the time these experiments were conducted, in the spring of 2023, there were no publicly available APIs for LLM-powered search tools such as Bing Chat or Google Bard/Gemini, and GPT 3.5 was the most advanced OpenAI model accessible through an API.

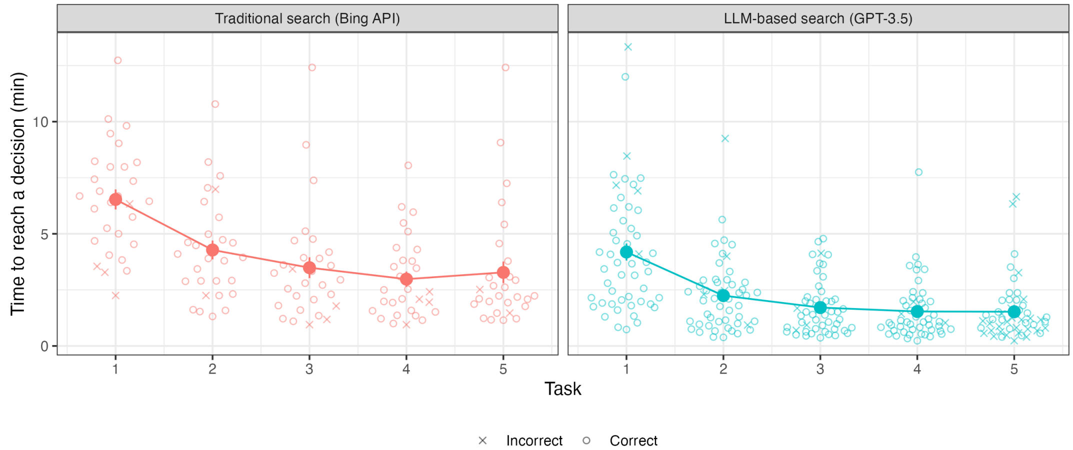



**(a) The main task interface. Participants are asked to choose between two vehicles. Instructions on the scenario and the metric of interest to make a decision are provided on the left, while the search tool is on the right. A notepad on the bottom left is also available for keeping track of found information.**

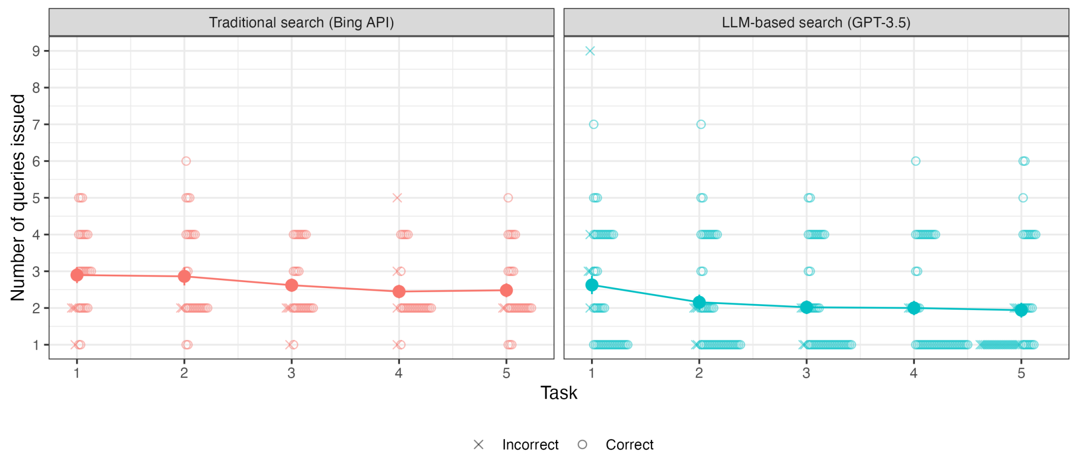



**(b) Search tool response interface for both conditions in this experiment: the experimental AI-powered search tool (left) and the experimental search engine (right).**

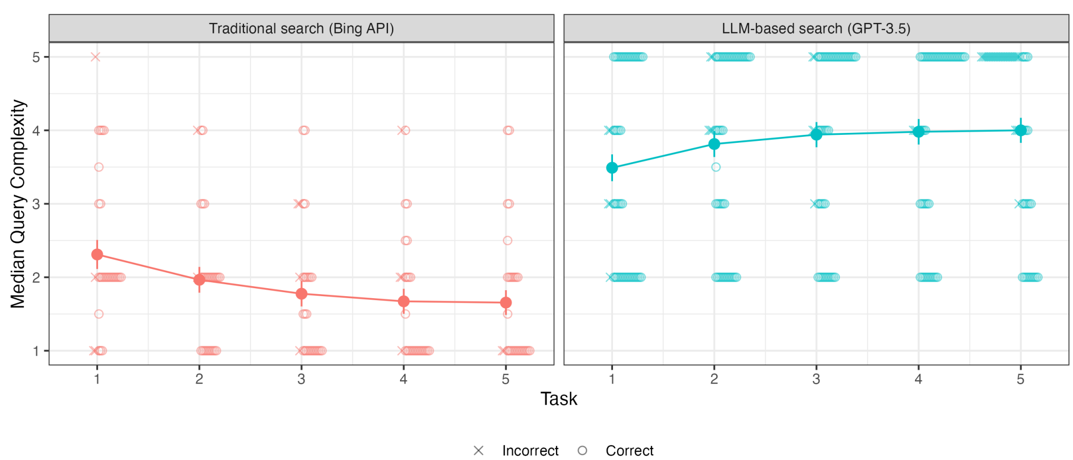
**Figure 1: Screenshots of the interface for Experiment 1.**




(a) Time to reach a decision by condition and task.




(b) Number of queries issued by condition and task.




(c) Complexity of queries issued by condition and task.

**Figure 2: Experiment 1: Efficiency and complexity results. Across all three figures each point represents one participant, with o's and x's indicating correct and incorrect responses (offset slightly on the x-axis for visibility), respectively. Solid points show the mean outcome in each condition across tasks, with error bars showing one standard error above and below the mean.**

Consistent with participants taking less time to answer with the LLM-based tool, participants issued fewer queries with the LLM-based tool as well, as shown in Figure 2b. With the LLM-based tool, most participants issued one query in all the tasks, while with the search tool, two queries was the most common pattern. Interestingly, many participants in the traditional search condition navigated to product information or comparison pages that allowed them to get both measurements for both vehicles in fewer than four, simpler queries for one product and dimension at a time.

We tested this difference with a generalized linear model, using a Poisson link function to model the number of queries by participant id as a random effect, task number as a control, and condition as a fixed effect (`glmer: num_queries ~ (1|worker_id) + as.factor(task_num) + condition, family = poisson`). The model revealed a modest but statistically significant main effect of condition on the number of searches (Estimate = -0.26, SE = 0.12, z = -2.244, p = 0.02). The estimated average number of queries made was 2.5 (95% CI [2.1, 3.0]) for the traditional search condition and 1.9 (95% CI [1.7, 2.2]) for the LLM-based search condition.

*Complexity.* While participants took less time and issued fewer queries in the LLM-based search condition, they made up for fewer queries by asking more complex queries. We average the complexity of each person by task in Figure 2c, where complexity is a number between 1 and 5 representing the number of unique elements of interest noted in the query. This could include 0, 1, or 2 products, 0, 1, or 2 dimensions, and 0 or 1 math question for the ratio of cargo space to length. Average complexity for the first task using traditional search was above 2 and decreased over rounds, while average complexity for participants with LLM-based search started above 3 and increased over rounds. Most of the gains, in both conditions, are between the 1st and 3rd task. Most LLM-based searches have complexity of either 2 or 5, with comparatively few at 3 or 4. Similar to the telemetry data noted in Table 1, a surprising amount of traditional search is of complexity 1 (i.e., a single product and no dimensions). While it takes a lot of time and many queries, these participants almost always make correct final decisions.

We tested this difference with a generalized linear model, using a Poisson link function to model the complexity of queries by participant id as a random effect, task number as a control, and condition as a fixed effect (`glmer: complexity ~ (1|worker_id) + as.factor(task_num) + condition, family = poisson`). The model revealed a statistically significant main effect of condition on the complexity of queries (Estimate = 0.65, SE = 0.09, z = 7.38, p < 0.001). The estimated average complexity of queries made was 1.8 (95% CI [1.6, 2.1]) for the traditional search condition and 3.4 (95% CI [3.1,3.8]) for the LLM-based search condition.

*Accuracy.* Figure 3 shows accuracy by task. For the first four easy tasks (comparisons between 8 popular, randomly-paired SUV models) accuracy was comparable between the two conditions, despite the traditional search users spending more time and issuing more queries to answer the questions. On the final task, designed to be difficult, (i. e., one where the LLM tends to err), participants' accuracy drops greatly due to mistakes in the LLM's responses, as the LLM tends to return the wrong cargo space for the Toyota 4Runner.

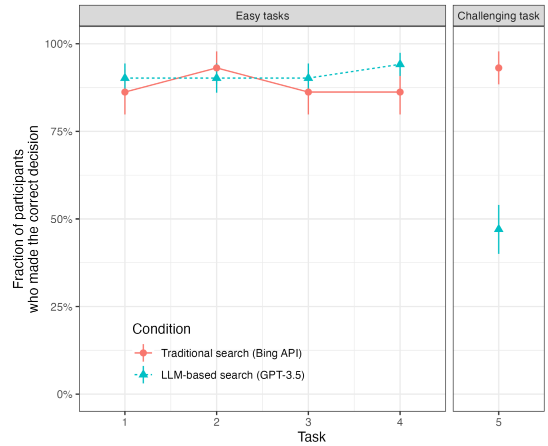



**Figure 3: Experiment 1: Accuracy by condition. The first four tasks are easy (comparisons between 8 popular SUV models), whereas the fifth is a comparison selected for which the LLM tends to err. Points represent means and error bars are plus or minus one standard error.**

To compare accuracy between conditions we fit a generalized linear model for the first four "easy" tasks. Specifically, we modeled whether participants made the right choice for each task, accounting for random effects by participant id, controls for the task, and a fixed effect for the condition they were assigned to (traditional vs. LLM-based search), with a logistic model (`glmer: is_correct_decision ~ (1|worker_id) + as.factor(task_num) + condition, family = binomial`). The fixed effects estimates revealed no significant effect of condition on the likelihood of making a correct decision for easy tasks (z = 0.99, p = 0.33). The estimated probabilities of making a correct decision, averaged over easy tasks, were 92.3% (95% CI [83%, 97%]) for the traditional search condition and 95.3% (95% CI [89%, 98%]) for the LLM-based search condition.

We used a separate generalized linear model to investigate accuracy in the final task, which was constructed to be challenging for the LLM. Specifically, we fit a logistic model with a fixed effect for the condition to predict whether participants made the right choice for this task (glmer: is_correct_decision ~ condition, family = binomial). The fixed effect estimate showed that the LLM-based search condition had a significant negative effect on the likelihood of making a correct decision compared to the traditional search condition (Estimate = -2.72, SE = 0.79, z = -3.46, p < .001). The estimated probabilities of making a correct decision were 93% (SE = 5%, 95% CI [76%, 98%]) for the traditional search condition and 47% (SE = 7%, 95% CI [34%, 61%]) for the LLM-based search condition.

The previous figures hint at what happened in the final task: participants who issued more complex queries were significantly more likely to select incorrect answers due to inaccurate LLM responses. Among the 51 participants in the LLM-based search condition, 30 submitted only one query. Of these, 23 received incorrect responses from the LLM and subsequently chose the wrong answer, while seven received accurate responses and selected the correct answer. Notably, most participants in the former group copied and pasted

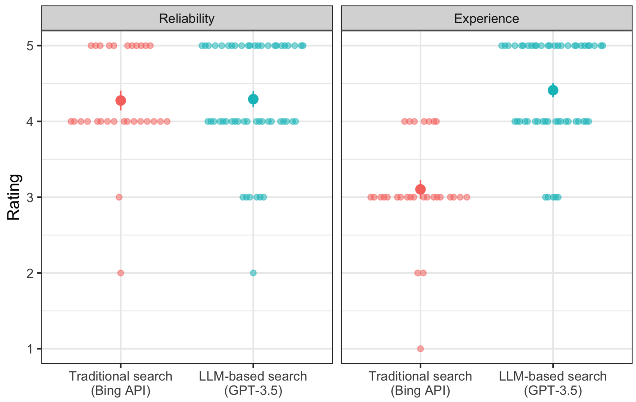



**Figure 4: Experiment 1: Results on user perceptions. Each smaller point represents one participant's response, the larger points show the mean by condition and error bars are plus or minus one standard error.**

the central question into the query box, whereas those in the latter group tended to rewrite the question in their own words.

Another 10 participants submitted exactly two queries. Six received accurate responses and chose correctly, while four received incorrect responses and made the wrong selection. As with the one-query group, their success or failure directly correlated with the accuracy of the LLM's answers.

Finally, 11 participants submitted four or more queries, all of which produced accurate LLM responses. Consequently, every participant in this group selected the correct answer. Interestingly, no participant attempted to refine their query or re-query a specific product or dimension after encountering an incorrect response.

*User experience and perceived reliability.* In the survey at the end of the experiment we asked participants to rate the reliability of the results they were shown and their overall search experience, both on 5 point Likert scales, with 1 being the worst and 5 being the best. As seen in Figure 4, perceived reliability was similar between conditions and overall quite high (4 or 5), despite the fact that many participants assigned to the LLM-based tool were exposed to incorrect information in the last task. We find no statistically significant difference in participants' subjective ratings of the reliability of the results that they were shown ($t(62.03) = 0.11$, $p = 0.91$), suggesting that users in the LLM condition who saw unreliable information were unaware of the errors made in the LLM output. In terms of the overall experience, participants strongly preferred completing the experiment with the LLM-based tool (with average rating of 4.41) compared to traditional search (with an average rating of 3.10), a statistically significant difference ($t(58.00) = 8.38$, $p < .001$).

Finally, analyzing the free response text that participants gave on the pros and cons of the LLM-based search tool revealed interesting insights. Overall, participants thought that the LLM-based search tool provided precise answers to specific queries, which was highly valued by participants. Additionally, they felt that the tool's quick responses were a major advantage, and its ease of use was appreciated for its simplicity and straightforwardness. Those who graded the tool as having high reliability noted the need for very specific search terms to achieve accurate results, while those who rated the tool with lower reliability expressed concerns about the lack of sources or citations, making it difficult to verify the reliability of the information. This divergence in feedback underscored the importance of both prompt specificity and source transparency in the tool's effectiveness.

## 5 Experiment 2

In the previous experiment we saw that while LLM-based search helped participants arrive at decisions faster, these decisions were not always of the same quality. Specifically, when LLM responses contained inaccurate information, it was difficult for participants to spot these mistakes due to a lack of cues about the veracity of the information they were shown. We designed our second experiment to investigate how people react to explicit cues that convey confidence in the responses generated by the LLM, and how this affects their decision making.

This experiment was a three condition, between-subjects design where all participants were assigned to use the same LLM-based search tool that responded with the same text to a given query, since it was again configured at zero temperature. The only thing that varied between conditions was how the numerical measurements in the responses were displayed visually via color coding. In the control condition, participants saw answers similar to those shown in Experiment 1—plain text without any cues about the veracity of measurements in the response. In each of the two treatment conditions, participants saw confidence-based color highlighting for numerical measurements contained in responses. As depicted in Figure 5, the "High + low confidence" condition showed green highlighting for "high confidence" measurements and red for "low confidence" ones, whereas the "Low confidence only" condition showed red highlighting for "low confidence" measurements only. The highlighting of each measurement was based on the token generation probabilities provided by GPT-3, with a generation probability of less than or equal to 50% displayed as a red highlight and greater than 50% displayed as a green highlight.[2,3]

The procedure was nearly identical to Experiment 1. Participants completed a sequence of three decision tasks comparing pairs of SUVs on the same criteria as in our first experiment (the ratio of total cargo space to total length). And, as in the first experiment, all but the last task were "easy" for the LLM in that there was a high likelihood of it returning correct information with high confidence, whereas the third task was once again "challenging" for the LLM and likely to contain inaccurate information, but with low confidence. We achieved this with GPT-3 by pre-prompting the model with ground truth measurements for the vehicles involved in each task *on everything except the first query of the third task.* This meant that the first and second tasks largely returned accurate information with

[2]We used GPT-3 in this experiment in order to have access to token probabilities, not available with GPT-3.5, which was used in the first experiment.
[3]Specifically, for measurements greater than 1, we used the token probability for the whole number token only (to the left of the decimal), whereas for measurements less than 1 we used the token probability for the decimal token only (to the right of the decimal. For example, for "47.2" the token probability for "47" is used, whereas for "0.248" the token probability for "248" is used. We piloted other highlighting schemes with more than two colors, but found these to be less effective than using just one or two colors.

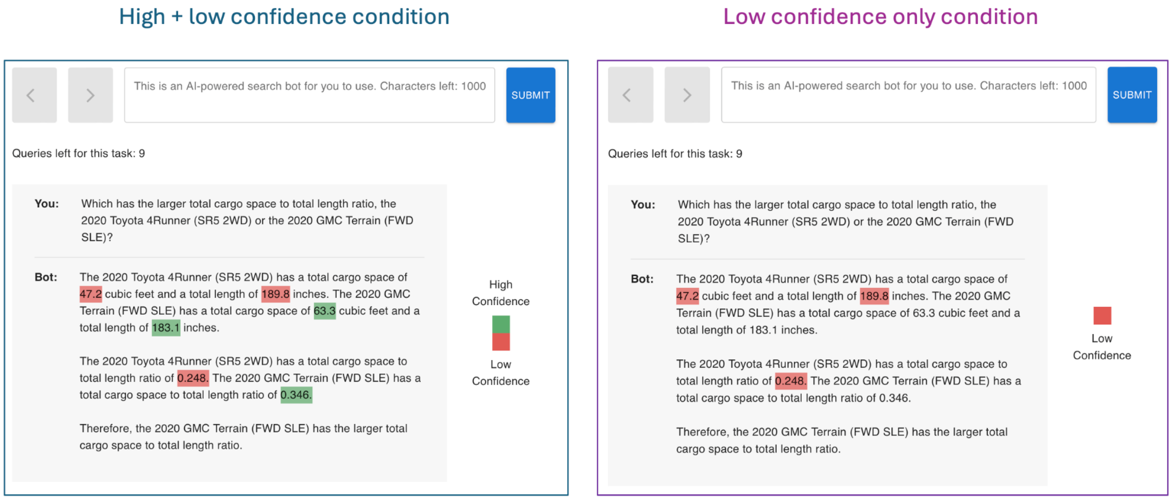



**Figure 5: The two treatments tested in Experiment 2: highlighting of both low and high confidence measurements (left) and only low confidence measurements (right). There was an additional control condition in which no highlighting was shown, mirroring Experiment 1.**

high confidence, but the first query of the third task often contained mistakes that were highlighted as low confidence. So if participants issued queries of the form "Which has the larger total cargo space to length ratio the 2020 Toyota 4Runner or the the 2020 GMC Terrain" on the first query of the third task, those in the treatment conditions would see cues about potentially unreliable information in the LLM response. The key question in this experiment was whether participants in the treatment conditions would take note of these low confidence cues and issue subsequent queries to double check the information they were shown.

We recruited 120 U.S. based participants from Amazon Mechanical Turk from a vetted pool of high-effort workers. For qualifications, we required at least 2,500 HITs approved with a 99% minimum approval rate. Participants were paid $5 for completing the experiment, with no performance bonuses. Similar to Experiment 1, we did not collect any demographic information from participants.

## 5.1 Results for Experiment 2

As in our first experiment, we analyzed efficiency, accuracy, and perceived experience across all conditions, but in this experiment we compare the three different treatments of confidence highlighting in LLM-based search instead of contrasting LLM-based search with traditional search.[4] For brevity we include only top-level results on the accuracy and perceived experience here, with the remaining results presented in Appendix C.

*Accuracy.* As in our first experiment, for the easy tasks (tasks 1 and 2) where the LLM provided largely reliable information with high confidence, accuracy was comparable between all three conditions, and quite high (Figure 6). However, for the challenging task

[4]Of note, results are not directly comparable between the two experiments because Experiment 1 used GPT-3.5 whereas Experiment 2 used GPT-3 due to the need for token probabilities.

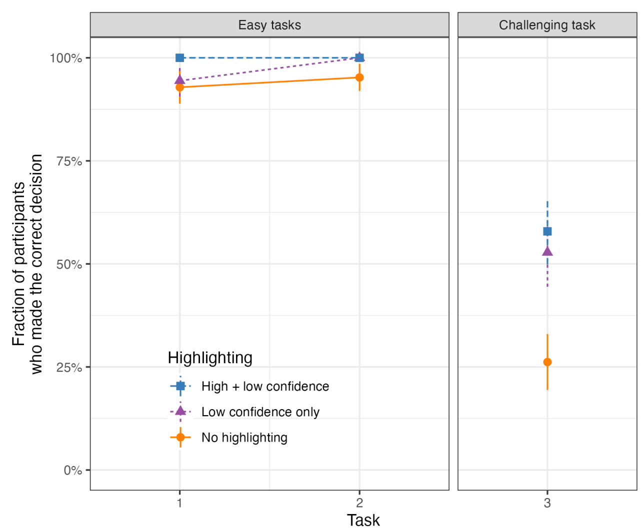



**Figure 6: Experiment 2: Accuracy by condition. The first two tasks are easy, whereas the third is a comparison selected for which the LLM tends to err. Points represent means and error bars are plus or minus one standard error.**

(task 3) where the LLM provided less reliable information on the first query, we see a dramatic difference between conditions: while accuracy plummets to 26% in the control condition without any confidence highlighting, accuracy in each of the treatment conditions was substantially higher—58% for the high + low confidence condition ($t(74.47) = -2.98$, $p < 0.01$) and 53% for the low confidence only condition ($t(70.36) = -2.44$, $p = 0.02$). In this case, both showing

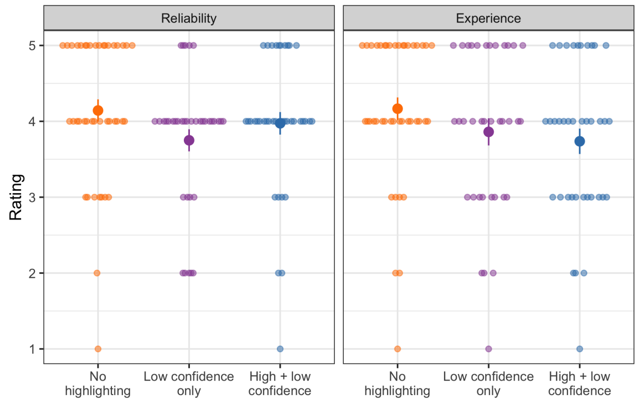



**Figure 7: Experiment 2: Results on user perceptions. Each smaller point represents one participant's response, the larger points show the mean by condition with error bars of plus or minus one standard error.**

high and low confidence cues and simply flagging low confidence information more than doubled accuracy in the decision task.

As shown in additional plots in the Appendix, the increased accuracy in the treatment conditions is largely due to participants issuing their initial query, seeing measurements flagged as low confidence, and issuing follow-up queries to double check the information they were shown. Whereas most participants in the control condition made a decision after one query, the majority of participants in the treatment conditions issued two or more queries, costing them some additional time, but more often leading to the correct decision.

To better understand the accuracy differences between conditions, we examined the rate of follow-up queries for each task. In tasks 2 and 3, 19 participants in the control condition initially issued a complex query comparing the two vehicles based on the ratio of both dimensions. Despite this, only 2 participants issued a meaningful follow-up query in either task, even though all 19 participants got the correct answer in task 2 but the wrong answer in task 3. In contrast, for participants in the treatment conditions, the number of participants making similar first queries increased from 24 in task 2 to 31 in task 3. More importantly, the number of meaningful follow-up queries rose from 5 to 15. This means that participants in the treatment conditions showed a slightly higher rate of follow-up queries than the control group when they already had the correct answer, but their rate of follow-up queries was much higher when their initial answer was incorrect.

*User experience and perceived reliability.* Finally, by way of perceived reliability and search experience, we find that all three conditions were rated quite favorably, and we detected no systematic difference between them, as shown in Figure 7. The no highlighting condition had an average reliability rating of 4.1, whereas the low confidence only condition had an average rating of 3.8 and the high + low confidence condition had an average rating of 4.0. Neither difference between the treatment conditions and the control is statistically significant (low confidence vs. no highlighting: $t(75.62) = 1.90$, $p = 0.06$; high + low confidence vs. no highlighting: $t(77.70) = 0.81$, $p = 0.42$). Similarly, the average search experience rating was 4.2 for the no highlighting condition, 3.9 for the low confidence only condition, and 3.7 for the high + low confidence condition. Despite a directional trend in the estimated means, neither difference is statistically significant (low confidence vs. no highlighting: $t(71.00) = 1.31$, $p = 0.19$; high + low confidence vs. no highlighting: $t(75.75) = 1.92$, $p = 0.06$).

Finally, we examined participants' free-text responses about the pros and cons of the LLM-based search tool in this experiment. Participants using the tool without confidence indicators (the same version as in Experiment 1) gave feedback consistent with those in the earlier experiment. They praised its ease of use, speed, ability to deliver precise information, and highlighted the importance of crafting detailed prompts to receive reliable responses. Those using the tool with confidence indicators similarly appreciated its efficiency in simplifying information retrieval and valued the confidence ratings as a way to enhance trust in the provided answers. However, they were more likely to express concerns about the lack of sources for verifying accuracy. This difference plausibly reflects how the confidence indicators drew their attention to questions of reliability, encouraging deeper consideration of the tool's limitations.

## 6 Discussion and Conclusion

In this work we investigated how LLM-based enhancements in search tools affect efficiency (time, number of queries, and query complexity), accuracy, user experience, and the ability to detect errors in a consumer research task. To obtain these measures, we developed an experimental platform that, holding all else constant, enables the random assignment of participants to either traditional or LLM-based search and keeps detailed records of their interactions.

In terms of efficiency, in our first experiment we found that participants who used the LLM-based search tool were able to complete tasks in roughly half the time compared to those who used a traditional search engine. In addition, we observed a slight reduction in the number of queries issued, accompanied by a significant increase in query complexity. Put simply, LLM-based search allowed people to reach decisions more rapidly and with fewer steps by issuing queries and receiving responses that more directly addressed the decisions at hand. These observed improvements in efficiency were accompanied by significant increases in favorable assessments of the LLM-based search tool based on participants' self-reports of their overall experience.

Concerning decision quality, in our first experiment we found comparable accuracy between conditions for easy tasks, but a significant drop in accuracy for those using the LLM when it erred, with almost half of the participants in this condition making an incorrect decision for the final task. In contrast, the vast majority of participants using traditional search made a correct decision for the final task. In investigating this drop in accuracy, we found that, without appropriate confidence cues, participants using LLM-based search were overreliant on the tool, with the majority of people (60%) issuing just a single query before reaching a decision. Furthermore, based on the nearly identical subjective reliability ratings between conditions, it appears that participants who used the LLM-based search tool were unaware of errors in the responses they were shown.

In our second experiment we proposed and tested mitigations for this issue of overreliance. We examined whether providing people with visual cues—specifically, color-coded responses that reflected the model's confidence based on token generation probabilities—could improve accuracy. This approach was effective: compared to a control condition without any highlighting, accuracy increased significantly, and participants were more inclined to seek additional information when responses indicated low confidence. This suggests that while automatically generating calibrated signals of model confidence is technically challenging, clearly communicating such signals to users can effectively reduce overreliance as they research products and services with LLM-based tools.

From a design perspective, the results of our experiments offer a clear message: if we want to encourage people to think critically about the information presented to them, we need to give them cues that help them do so. Otherwise it is likely the case that they will take what they are shown at face value. Thankfully, however, even very simply cues such as coarse-grained highlighting of potentially unreliable information can help. This highlights a broader design principle: effective tools should not only deliver accurate information but also actively support users in questioning and verifying that information.

Two open questions are: how best to identify potential errors in LLM-based outputs and how best to convey confidence to users, enabling them to make informed decisions. In our experiments, we used token probabilities to identify potentially incorrect responses. In order to do that, we had to move from GPT-3.5 to GPT-3.0, because token probabilities were not available for GPT-3.5 and later models. Token probabilities are not necessarily calibrated with respect to correctness, which can be exacerbated when models are updated based on human preference data [37]. In ongoing research, we are investigating more reliable ways to identify incorrect information in LLM output and to convey the various types of errors that LLMs might make, and how to best provide cues about potential errors to users.

While this investigation found many benefits of LLM-based search, it also uncovered an unanticipated strength of traditional search. In practice, it is rare for people to query more than one product and one dimension at a time, despite how common it is for users to compare products on multiple dimensions [6, 20, 40]. In our first experiment, participants in all conditions were effectively encouraged to try more complex searches. Surprisingly, participants in the traditional search condition found correct answers in fewer than four queries on average; sometimes they found them in just one query. In our experiments this paid off because the traditional search tool often directed people to pre-generated product comparison web pages. Therefore, it is possible that the introduction of LLM-based search norms may encourage people to issue more complex queries to traditional search engines, improving their efficiency with these existing tools.

As to limitations, we explored a specific search domain and a simplified decision task. Specifically, we asked participants to make choices among pre-defined pairs of vehicles. While this allowed us to construct well-defined tasks with clear and correct answers, it also represents just a portion of the broader search process. In practice, search is often an open-ended and iterative journey where people conduct research to learn about a given domain and refine what they are searching for as they go. While we leave more open-ended studies for future work, we anticipate that LLM-based search could offer even greater advantages in such scenarios, as LLMs are better equipped than traditional search to respond to open-ended questions and to synthesize and compare resulting information. Future work could also explore other aspects of the broader search process. We focused on searches involving the purchase of high cost durable goods, but we expect that different search scenarios will be affected in different ways. For example, while users may quickly move to more complex queries when doing research on purchasing a car, they may continue to use simpler searches in other product categories.

Furthermore, we expect users to respond differently to cues about the correctness of information under different conditions. For example, we observed that when the LLM-based tool provided signals of its own low confidence in certain measurements, it led people to carry out more searches for additional information in a car buying scenario. However, users may be inclined to accept such pieces of information in other scenarios if they are more interested in a range or approximate values than in exact numbers (for example when searching for the number of calories in a specific type of food). Another question for future research is the degree to which members of various demographic groups (e.g.,those with higher or lower levels of educational attainment) would adaptively alter their behavior in response to confidence signals. If differences exist, it might suggest tailoring different signals to different audiences.

We also designed our LLM-based search tool around a non-conversational version of GPT for a tight experimental contrast with traditional search. A natural avenue for future investigation would be to include conversational capabilities, and to explore newly-available tools such as Bing Chat, Google Gemini, and the latest ChatGPT models that blend traditional search and LLMs by issuing traditional queries against a web index and using LLMs to summarize the results, including external hyperlinks for further information [24]. Future work could also compare different methods for reducing overreliance. For example, the confidence-based highlighting studied here could be evaluated against alternative approaches for communicating uncertainty, such as qualifying statements, numerical confidence scores, or providing references to source materials via hyperlinks.[5] While our work establishes confidence-based highlighting as one method to combat overreliance, this is not to say it is the most effective method for doing so, or should be the only one deployed in practice.

A final potential limitation of this study is that, while our design encouraged participants to engage thoughtfully with the tasks, we cannot completely rule out the possibility that some participants failed to invest sufficient effort or misinterpreted the task as requiring them to echo the tool's answers rather than using the tool to make correct choices. While this is theoretically possible, random assignment to conditions should ensure that any variations in motivation or task interpretation are evenly distributed across

[5]At the time that these experiments were conducted, there were no public APIs that incorporated retrieval augmented generation based approaches and so we had no way to provide references or citations for the LLM's responses.

conditions. The total degree of overreliance might vary with the incentives offered, but we expect the differences between conditions to generalize.

Although incentives for accuracy might have further increased effort, such approaches could also backfire by encouraging participants to disregard instructions and seek information outside the experiment. Research has shown that incentives can sometimes negatively impact participants' voluntary cooperation or the truthfulness of their responses [14, 44].

Despite these considerations, our findings provide robust insights into how participants interact with LLM-based and traditional search tools under controlled conditions. The significant improvement in accuracy with confidence-based highlighting in Experiment 2, the high and consistent subjective reliability ratings across conditions, and the thoughtful qualitative feedback from participants all suggest meaningful engagement with the tasks. These results highlight the potential of LLM-based tools to improve decision-making when thoughtfully designed. Further studies could explore how varying incentive structures or real-world task contexts might influence behavior.

In sum, LLM-based information retrieval tools stand to permanently change how users search for information about products and services online. The studies presented here suggest that search efficiency and user satisfaction will likely increase, but overreliance may become more of a concern. We anticipate ongoing innovation and evaluation in methods of conveying uncertainty in AI responses so that they may be viewed with an appropriate level of confidence. Uncertainty in AI answers, and in the world for that matter, can never be eliminated, but effective means of communicating it can augment people's cognition and decision making. We view this as a key challenge for the field of human-computer interaction in a world that is increasingly relying on AI tools that are themselves not 100% reliable.

## Acknowledgments

We thank Jiawei Liu for the initial development of the web framework used in this experiment. We thank Mark Whiting and Duncan Watts for assistance in recruiting participants in the second experiment. We thank the audience at University of Pennsylvania's "Large Language Models: Behavioral Science Meets Computer Science" workshop on May 19, 2023, and Microsoft Research's "AI, Cognition, and the Economy workshop" on October 12, 2023. We also thank Susan Dumais for her invaluable feedback.

## References

[1] Eugene Agichtein, Steve Lawrence, and Luis Gravano. 2004. Learning to find answers to questions on the web. *ACM Transactions on Internet Technology (TOIT)* 4, 2 (2004), 129–162.

[2] Marcia J Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online review* 13, 5 (1989), 407–424. Publisher: MCB UP Ltd.

[3] Michael Bendersky and W. Bruce Croft. 2008. Discovering key concepts in verbose queries *(SIGIR '08)*. Association for Computing Machinery, New York, NY, USA, 491–498. https://doi.org/10.1145/1390334.1390419

[4] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 185–194.

[5] Sergey Brin. 1998. The PageRank citation ranking: bringing order to the web. *Proceedings of ASIS, 1998* 98 (1998), 161–172.

[6] Bart J Bronnenberg, Jun B Kim, and Carl F Mela. 2016. Zooming in on choice: How do consumers search for cameras online? *Marketing Science* 35, 5 (2016), 693–712.

[7] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. 2023. *Generative AI at work*. Technical Report. National Bureau of Economic Research.

[8] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[9] Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[10] Lydia B Chilton and Jaime Teevan. 2011. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*. 27–36.

[11] Fabrizio Dell'Acqua, Edward McFowland, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* 24-013 (2023).

[12] Susan Dumais, Edward Cutrell, and Hao Chen. 2001. Optimizing search by showing results in context. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 277–284.

[13] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*. 406–414.

[14] Simon Gächter, Esther Kaiser, and Manfred Königstein. 2024. Incentive contracts crowd out voluntary cooperation: Evidence from gift-exchange experiments. (2024).

[15] Manish Gupta and Michael Bendersky. 2015. Information Retrieval with Verbose Queries *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 1121–1124.

[16] Ahmed Hassan Awadallah, Ryen W White, Patrick Pantel, Susan T Dumais, and Yi-Min Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. 829–838.

[17] Bernard J Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management* 36, 2 (2000), 207–227.

[18] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 607–616.

[19] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[20] Jun B Kim, Paulo Albuquerque, and Bart J Bronnenberg. 2011. Mapping online consumer search. *Journal of Marketing Research* 48, 1 (2011), 13–27.

[21] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.

[22] Steve Lawrence. 2000. Context in Web Search. *IEEE Data Engineering Bulletin* 23, 3 (2000), 25–32.

[23] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2024. Hill: A hallucination identifier for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[25] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=8s8K2UZGTZ

[26] Jiqun Liu. 2021. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management* 58, 3 (2021), 102522. Publisher: Elsevier.

[27] Jia Liu and Olivier Toubia. 2018. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science* 37, 6 (2018), 930–952.

[28] Jia Liu and Olivier Toubia. 2020. Search query formation by strategic consumers. *Quantitative Marketing and Economics* 18 (2020), 155–194.

[29] Jia Liu, Olivier Toubia, and Shawndra Hill. 2021. Content-based model of web search behavior: An application to TV show search. *Management Science* 67, 10 (2021), 6378–6398.

[30] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848* (2023).

[31] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[32] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661* (2020).

[33] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 1055–1058.

[34] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[35] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192.

[36] Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1551–1560.

[37] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[38] Samir Passi, Shipi Dhanorkar, and Mihaela Vorvoreanu. 2024. *Appropriate reliance on Generative AI: Research synthesis*. Technical Report MSR-TR-2024-7. Microsoft. https://www.microsoft.com/en-us/research/publication/appropriate-reliance-on-generative-ai-research-synthesis/

[39] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature Review*. Technical Report MSR-TR-2022-12. Microsoft. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[40] John W Payne, James R Bettman, and Eric J Johnson. 1993. *The Adaptive Decision Maker*. Cambridge university press.

[41] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590* (2023).

[42] Girish N Punj and Richard Staelin. 1983. A model of consumer information search behavior for new automobiles. *Journal of Consumer Research* 9, 4 (1983), 366–380.

[43] Abigail Sellen and Eric Horvitz. 2024. The rise of the AI Co-Pilot: Lessons for design from aviation and beyond. *Commun. ACM* 67, 7 (2024), 18–23.

[44] Kathryn Sharpe Wessling, Joel Huber, and Oded Netzer. 2017. MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research* 44, 1 (2017), 211–230.

[45] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, Vol. 33. ACM New York, NY, USA, 6–12.

[46] StatCounter. 2023. *Search engine market share worldwide*. https://gs.statcounter.com/search-engine-market-share

[47] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–24.

[48] Emma van Inwegen, Zanele T Munyikwa, and John J Horton. 2023. *Algorithmic writing assistance on jobseekers' resumes increases hires*. Technical Report. National Bureau of Economic Research.

[49] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q Vera Liao, and Jennifer Wortman Vaughan. 2024. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *TOCHI 2024: ACM Transactions on Computer-Human Interaction* (2024).

[50] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[51] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do Large Language Models Know What They Don't Know?. In *Findings of the Association for Computational Linguistics: ACL 2023*. 8653–8665.

[52] Albert Ziegler, Eirini Kalliamvakou, X Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity assessment of neural code completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 21–29.

**Appendix**

## A Traditional Search Activity

Table 1 provides a breakdown of Bing searches for top SUVs in terms of the number of vehicles and dimensions in the first half of 2022.

| Number of Vehicles | Number of Dimensions | Percent of Queries | Percent of Queries (w/ 1+ dimension) |
|---|---|---|---|
| 1 | 0 | 63.00% | – |
| 1 | 1 | 25.18% | 68.40% |
| 1 | 2 | 9.22% | 25.05% |
| 1 | 3 | 2.00% | 5.43% |
| 1 | 4 | 0.23% | 0.63% |
| 1 | 5 | 0.01% | 0.01% |
| 1 | 6 | 0.00% | 0.00% |
| 2 | 0 | 0.18% | – |
| 2 | 1 | 0.10% | 0.26% |
| 2 | 2 | 0.07% | 0.18% |
| 2 | 3 | 0.01% | 0.04% |
| 2 | 4 | 0.00% | 0.00% |
| 3 | 0 | 0.00% | – |
| 3 | 1 | 0.00% | 0.00% |

**Table 1: Table of number of products and dimensions in all searches for top 25 SUV in 2022. Starting with a list of the top 25 SUV by sales in the first half of 2022 we looked at every 2022 Bing search that included these 25 SUVs and and the top 10 most queried dimensions (e.g., cargo space, length, etc). Most queries mention only one vehicle. If a dimension is mentioned, most queries mention only one dimension.**

## B Experiment 1

### B.1 Tutorial

Before starting the first task, participants in each condition were given a short tutorial on what to expect from the search tool they would be using. Figure 8 shows the tutorial for the traditional search condition and Figure 9 shows the tutorial for LLM-based search.

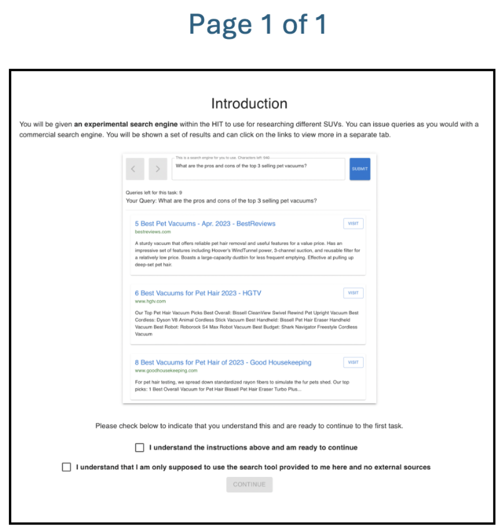



**Figure 8: The tutorial for participants who were in the traditional search condition (Experiment 1).**

### B.2 Speed and accuracy jointly

Speed and accuracy are both desirable for search engine users. Figure 10 plots them against each other. The upper left corner of each panel represents the best performance, that is, the most correct answers in the least amount of time. To facilitate seeing patterns despite overplotting, a density was fit to the responses to create a heat map. The high density areas in both panels show the participants with the LLM tools in a favorable position near the upper left. They have less variance in time taken but more variance in accuracy, mostly owing to the additional item designed to be difficult. Performance on this item is marked with an x showing that the vast majority of participants who did not score all the questions correctly made an error on this item.

## C Experiment 2

### C.1 Efficiency

As in our first experiment, across conditions we see a learning effect where participants take less time to reach a decision on the second task compared to the first (Figure 11). Using a similar linear mixed model as in Experiment 1 to model log task duration on the easy tasks, we find that on average across all conditions, participants take 3.3 minutes (95% CI [2.9 minutes, 3.7 minutes]) to complete the first task, but only 1.8 minutes (95% CI [1.6 minutes, 2.0 minutes]) to complete the second task. Averaged over both of these tasks, we find that participants in the treatment conditions were slightly slower than

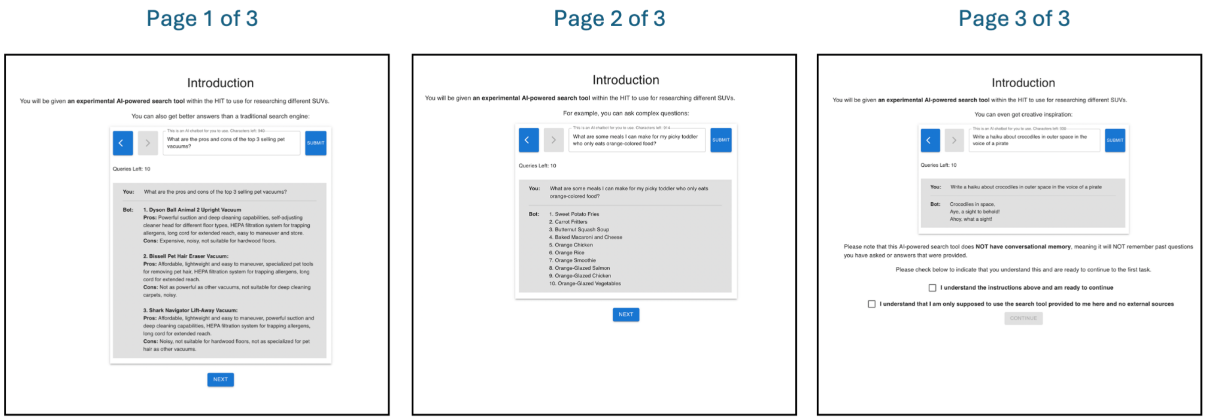



**Figure 9: The tutorial for participants who were in the LLM-based search condition (Experiment 1).**

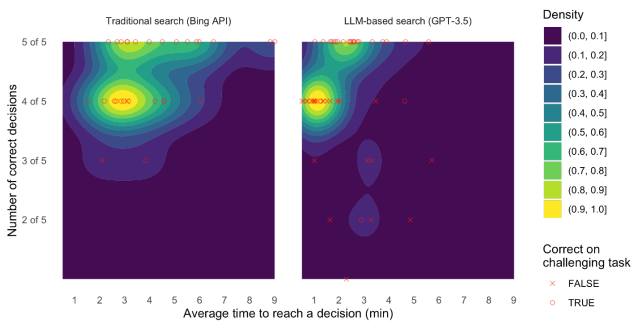



**Figure 10: Joint view of speed and accuracy (Experiment 1). Each point represents the data from one participant over five questions. Points are represented with an "o" if they got the challenging question correct and a "x" if they failed to.**

those in the control overall, with a statistically significant difference for high + low confidence highlighting compared to no highlighting ($t(113) = 2.09$, $p = 0.04$) but no evidence of a systematic difference for low confidence highlighting only ($t(113) = 0.63$, $p = 0.53$). On the third task, where participants encounter potentially unreliable information, we see an increase in time to decision for the two treatment conditions that highlight potentially unreliable information, but no such increase for the control without confidence highlighting (low confidence only vs. no highlighting: $t(72.79) = -2.53$, $p = 0.01$; high + low confidence vs. no highlighting: $t(72.63) = -3.70$, $p < 0.001$).

Analyzing the number of queries using a similar linear mixed model as in Experiment 1, we find no evidence of systematic differences in the number of queries issued in the first two easy tasks across conditions, with participants issuing 2.3, 2.7, and 2.7 queries per task on average for the no highlighting, low confidence only, and high + low confidence conditions, respectively. However, in the third task we see substantial increases in the number of queries for the two treatment conditions compared to the control (low confidence only vs. no highlighting: 3.0 vs 2.2 queries on average, $t(70.97) = -2.00$, $p = 0.05$; high + low confidence vs. no highlighting: 3.6 vs 2.2 queries on average, $t(73.21) = -3.29$, $p = 0.002$). This is visually apparent in Figure 12, as depicted in uptick in queries for the middle and right panels compared to the left panel.

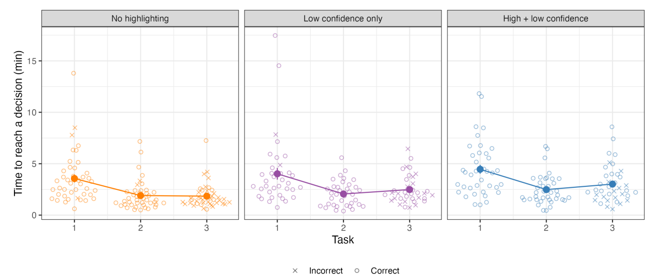



**Figure 11: Time to reach a decision in Experiment 2 by condition and task. Each point represents one participant's number of queries for the task, with o's and x's indicating correct and incorrect responses, respectively.**

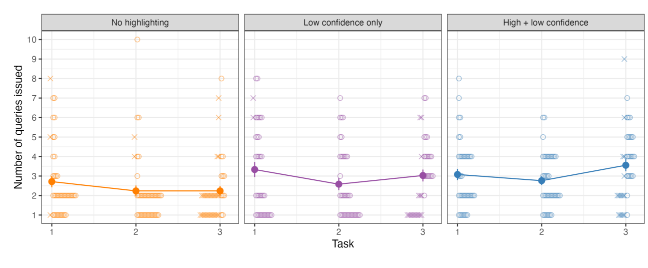



**Figure 12: Number of queries issued in Experiment 2 by condition and task. Each point represents one participant's number of queries for the task, with o's and x's indicating correct and incorrect responses, respectively.**