Master Thesis

Comparing User Behavior in Information Retrieval Using Traditional Search Engines and LLMs

Boleslav Khodakov

Delft University of Technology

TU Delft

Master Thesis

Comparing User Behavior in Information Retrieval Using Traditional Search Engines and LLMs

by

Boleslav Khodakov

to obtain the degree of Master of Science
in Management of Technology
at the Delft University of Technology

Thesis committee: Dr. Laurens Rook, 1st supervisor
Dr. Iulia Lefter, chair and 2nd supervisor

Cover: "a person using a laptop" by Árpád Czapp on unsplash.com
Style: TU Delft Report Style, with modifications by Daan Zwaneveld (modified)

An electronic version of this thesis is available at https://repository.tudelft.nl/

TU Delft

Preface

Artificial intelligence has been a recurring theme in my academic work since my bachelor thesis. Over the past years, I have become increasingly fascinated by the speed at which AI has developed from a specialized research topic into a technology that influences everyday workflows, decision making, and information retrieval. What once seemed distant and experimental has rapidly become practical, accessible, and deeply embedded in how people search for and process digital content.

This thesis was written during a time in which Large Language Models are changing the relationship between humans and digital information. They offer new possibilities for making work more efficient, reducing friction in complex tasks, and supporting people in finding answers more naturally. At the same time, they raise important questions about accuracy, trust, satisfaction, and over-reliance. These questions are central to this thesis, which compares user behavior in traditional search engines and LLM-based search environments.

With this thesis, I hope to make a modest contribution to that broader discussion. Understanding how users interact with AI-based tools is necessary if we want these systems to become not only more powerful, but also more reliable, transparent, and genuinely useful.

This study would not have been possible without the support I received from my supervisors, family, and friends.

First, I would like to sincerely thank my first supervisor, Professor Laurens Rook, for his guidance throughout every stage of this master thesis. From the initial research idea to the final writing phase, his advice, critical reflections, and continued support helped shape this thesis into its final form.

I would also like to thank my second supervisor, Professor Iulia Lefter, for her thoughtful reviews and constructive criticism during important stages of the thesis process. Her feedback helped me refine my arguments and look at my research with a sharper analytical perspective.

Furthermore, I would like to thank Professor Markus Zanker from the Free University of Bozen-Bolzano for his valuable contributions during the development stage of the experiment.

Last, but certainly not least, I would like to give a heartfelt thank you to my friends and family, who stood by me not only during this thesis, but throughout my studies as a whole. Their unwavering support was a ray of sunshine during difficult times and a source of even greater happiness during good times.

Executive Summary

With the introduction of ChatGPT™ in 2022, Large Language Models (LLMs) have changed the way people interact with digital information. From writing support to image generation and business reporting, LLMs have become useful in many workflows. One area where they may be especially relevant is Context-Aware Recommender Systems (CARS), which already support everyday recommendations for music, products, jobs, and other forms of information. Although research interest in CARS has recently declined, LLMs may offer a new opportunity by enabling more dynamic and conversational interactions between users and systems.

This is important because searching for information online is not always simple. Information is spread across many sources, differs in quality, and often requires users to decide when they have searched enough. Some users continue searching until they believe they have found the best possible option; these users are known as maximizers. Others stop once they find an option that is good enough; these users are known as satisficers. Since these differences may affect how people search and evaluate outcomes, this thesis compares not only the accuracy of different search tools, but also user satisfaction and the role of personality.

The goal of this thesis is to compare how LLM-based search and traditional search engines support information retrieval. Specifically, the study examines whether the search tool affects result accuracy and user satisfaction, and whether these effects are moderated by two components of maximization: maximization goal and maximization strategy. Maximization goal refers to the desire to choose the best possible option, while maximization strategy refers to the tendency to search extensively before making a decision. The study hypothesized that LLM-based search would lead to lower accuracy than traditional search, but higher satisfaction. It was also expected that maximization goal and strategy would moderate these effects.

To test these hypotheses, a live experiment was conducted with participants recruited at TU Delft and Leiden University. Participants were randomly assigned to one of two search tools: DuckDuckGo™, representing a traditional search engine, or OpenRouter™ with Grok™ 4.3, representing an LLM-based search environment. Before completing the search tasks, participants filled in a personality questionnaire measuring maximization goal and maximization strategy. They then completed two independent tasks using their assigned tool. The first task required participants to identify a former TU Delft student and answer questions about the student's thesis and related academic publication. The second task required participants to identify apple conditions from images. After each task, participants submitted their answer and rated their satisfaction with the search process.

The results showed no significant difference in accuracy between the LLM and search-engine conditions. Therefore, the hypothesis that LLM-based search would lead to lower accuracy was not supported. However, accuracy was very low across both tasks, which makes this result difficult to interpret. The tasks likely created a floor effect, meaning that they were too difficult to clearly detect differences between the tools. Therefore, this

finding should not be interpreted as evidence that LLMs and search engines are equally accurate in general.

For satisfaction, the results were clearer. Participants using the LLM reported significantly higher satisfaction than participants using DuckDuckGo™. Descriptive behavioral results also showed that LLM users completed the tasks faster, entered fewer queries, and visited fewer external websites. Search engine users, in contrast, searched more broadly across websites and domains. This supports the idea that traditional search engines encourage navigation across multiple sources, whereas LLMs concentrate the search process within a single conversational interface.

The moderation analyses showed that neither maximization goal nor maximization strategy moderated the relationship between search tool and accuracy. In other words, users' maximization tendencies did not significantly change how accurately they performed with either DuckDuckGo™ or the LLM. For satisfaction, maximization goal also did not significantly moderate the effect of search tool. Maximization strategy, however, did moderate the relationship between search tool and satisfaction. The satisfaction advantage of the LLM was strongest among participants low in maximization strategy, but disappeared among participants high in maximization strategy. This suggests that users who do not naturally search extensively may benefit more from the guided structure of an LLM. Users high in maximization strategy may instead value comparison, visible alternatives, and control over the search process, which are more naturally supported by traditional search engines.

Overall, this thesis shows that LLMs can make information retrieval more satisfying, but that higher satisfaction does not automatically imply higher accuracy. The findings suggest that LLM-based systems should include verification mechanisms, such as source links, uncertainty indicators, or prompts encouraging users to check important outputs. They also suggest that future CARS and search platforms may benefit from adapting to users' decision-making styles. As LLMs become increasingly integrated into search and recommender systems, it is important to understand not only when these tools work, but also for whom they work best.

Contents

1. Introduction

A recent position paper highlighted a declining research interest in Context-Aware Recommender Systems (CARS), largely attributed to the rising prominence of Deep Learning models [42]. The paper argued that a next-generation of CARS should leverage AI-assistants driven by Large Language Models (LLMs). This proposition stemmed from the belief that CARS incorporating LLMs could foster a more dynamic, detailed, and transparent dialogue between humans and agents, ultimately enhancing recommendation quality. They note that, due to the rise of AI chatbots, people are starting to write their queries more explicitly again, which can be used for relevant data extraction. Simultaneously, the researchers warn of simply using implicit context, found in Deep Learning systems. [42] explain that compared to Machine Learning, where context is explicitly added by developers, Deep Learning has to guess what context is useful; its contextual quality therefore suffers.

The idea that recommendation systems can be enhanced through the inclusion of LLMs is getting increasingly popular [cf. 60]. However, for CARS to be successful, the information provided to the system must be of high quality. That is, incorrect information or withheld data risks leading to a less relevant recommendation. A classic example is users of AI-assistants not willing to disclose relevant personal information due to privacy concerns [cf. 41]. However, users may also be unaware of the correct ways of presenting information to LLMs [cf. 36]. There exist techniques of building a query for an LLM known as “prompt engineering”. However, users must explicitly learn about these techniques as they may not be obvious at first. It is exactly these tiny details, known as context, that can make (or break) someone’s search by catering it to their circumstances [2]. By incorporating context into their recommendations companies can enhance everything from music searches to job connections [2].

An old-school competitor to AI chatbots is the traditional search engine (e.g. Google™). Both tools can be used for information retrieval. Established literature has described this process in detail [38, 37]. Interestingly, some researchers argue that users forage information differently on LLMs compared to search engines [39]. Whereas search engines are argued to provide broader information with simpler queries, LLMs require more precise input. LLMs work worse for scattered information, and they limit the user to one single source (themselves).

Additionally, we suspect that a person’s personality may have an impact on how they search for information. Psychological literature explains that there exist two types of individuals - ”maximizers” and ”satisficers” [49, 46]. Satisficers seek solutions which are sufficient to solve their problem, whereas maximizers require the best possible solution for their problem [49, 46]. It has been shown that this personality trait impacts several behaviors - maximizers tend to be more risk averse and less happy with their solutions than their satisficing peers [25, 43, 6, 53]. For CARS, the results of personality type affecting outcomes are split. Some literature found little correlation [34, 22], while others observed differing behavior between the two groups such as ”choice overload” and deeper AI workflow integration [9, 45].

In this MSc thesis, we investigate several problems. The main research objective is to see whether the use of AI tools versus search engines leads to differences in accuracy and user satisfaction when it comes to information retrieval. We hypothesize that traditional search engines will provide more accurate results, yet users will find LLMs more satisfying to use.

Our second research objective focuses on whether people's personality affects the way they research information. In practice, we want to see whether personality moderates the relationship between search tool and accuracy or satisfaction. We hypothesize that users' maximization tendencies will moderate the relationship between search tool and both accuracy and satisfaction.

We conduct an experiment where participants receive two tasks about retrieving information while being assigned either a search engine or an LLM chatbot as a tool to help them out. They may complete the tasks only using this single tool. After each task, participants provide an answer to the questions from the task and rank their satisfaction with the search process. Moreover, before the experiment participants fill out a personality questionnaire for us to assess whether they are a maximizer or satisficer. Through recording user behavior during task completion (with recording software) we also collect other interesting individual differences in search practices, such as varying task completion speed, for a future study.

This thesis is structured as follows. In section 2 the background surrounding the proposed user study is outlined. Specific attention is given to how humans seek information, what Context-Aware Recommender Systems are, and maximization theory. Next, in section 3 we outline the present study. Our research questions and the hypotheses they lead to are established. In section 4 the search tasks used in this study are presented. Section 5 elaborates on the instruments used in the study, while section 6 outlines the experimental procedure and design. Our results are presented next in section 7. Lastly, section 8 concludes the study with the interpretation of the results, suggestions for future work, and a formal conclusion.

2. Background

2.1. Context-Aware Recommender Systems (CARS)

Recommender systems are important parts of the information retrieval process and create value for multiple stakeholders [21, 4]. [2] identifies three types of incorporating context: through explicit, latent, and dynamic context. Latent context is especially interesting as Deep Learning is a popular way to gather the hidden patterns in data. Netflix™, Spotify™, and LinkedIn™ are commercial examples that use contextual information in their products to enhance customer experience [2]. One such experience is information retrieval with the help of LLMs, for which the quality of context matters. [10] found that users often prefer items based on ratings from similar users, or their own past ratings on related items. Meanwhile, the researchers found ratings from social network connections not to be reliable. In [59] and [21] it was found that simple improvements in Machine Learning accuracy does not enhance user experience. [2] outlines that for context to be useful, it has to be relevant, representational and able to be captured.

[42] express fear that research of CARS is in decline. They speculate that there are two main reasons - the (addressable) ethical considerations and the rise of Large Language Models. [15] have tried bridging the LLM-context gap in their research. The researchers created an algorithm that explicitly asks the user for context. In the second stage of their "Conversational Prompt Engineering" algorithm, the user is specifically asked for feedback on example outputs. This feedback actively changes a prompt from being zero-shot (if the user were to ask one question) to being few-shot. Another study with personalization is [55]. Before interacting with an AI chatbot, users completed a pre-task questionnaire capturing task context and user perceptions, such as topic familiarity and expected complexity. This information was then incorporated into prompt templates through prompt engineering to enrich the interaction context. In addition to the questionnaire, the system provided supportive functions such as prompt suggestions and conversation explanations. The combination of these functions reduced cognitive load and helped users better manage expectations.

Conversely, search engines are, without any additional tweaks, an example of CARS. [2] provide "mobile information search" as a direct example of CARS. We may extend this definition to "web search engines". The authors outline a convenient example of a typical search enhanced by context: If a user searches for "nearby restaurants that are currently open", the user's mood, their location, and the time of day are great factors to refine a query. These contextual factors help the system estimate what information is useful in the user's current situation. This connects to the account of information sampling and foraging from [3], where decisions depend on contextual factors such as uncertainty and time horizon.

While they are the most crucial to this research, the differences in context handling are not the only factor differentiating LLMs from search engines. Whether it is their context, user experience, or architecture, researchers have begun comparing the two rivaling technologies head-to-head. These techniques will be presented in the next section.

2.2. Traditional Web Search Engines vs. Large Language Models

Currently on the web, there are two main ways of researching information. The traditional way is using a SERP[1]-based search engine. Examples include Google™, Bing™, and Yandex™. A query is written by the user which is then compared against an index [54]. The results, which are deemed most relevant by the engine's algorithm (e.g. Google™'s PageRank™), are then displayed as a list to the user.

Since 2022, another method for retrieving information is rapidly gaining popularity on the web. The initial launch of OpenAI™'s ChatGPT™ created the option to ask questions to chatbots which are powered by Large Language Models [17]. These LLMs are able to process a natural-language-based query instead of being keyword-based like traditional search engines [31]. After processing the query, the LLM generates a unique response which is displayed to the user in a chat-like window. To aid accuracy, many LLMs also have access to web search capabilities.

The rise of LLMs led to different usage patterns compared to search engines. In [50] these differences were evaluated. When researching topics, participants using an LLM spent roughly half the time of those using traditional search engines to get to the answer. However, when the LLM was erroneous, the accuracy dropped to 47%, while search engine users' accuracy was consistently above 92%. Yet, despite this significantly higher error rate, users were more satisfied with the LLM.

A second study also found time savings with the "Bing™ Chat" LLM, along with higher user satisfaction as opposed to using traditional Bing™ Search [27]. Worryingly, however, for 53% of Bing™ Chat searches users did not check any sources except the generated response itself. In both [50] and [27] the traditional search engine study group consistently referenced various links. This stark contrast may be an indication of user over-reliance on LLMs. It can be argued that people blindly trust a coherently written text, despite every LLM chatbot coming with a warning of hallucination.

Moreover, [44] suggest that any bias present in LLMs that it shares with the user may polarize said user further. When LLMs were biased to agreeing with the user, the study found a sudden increase in querying. A study analyzing crowdsourced interactions with Google™ Bard™ found that participants were also not aware of how to optimize their search queries [51]. For instance, there were surprisingly few prompt engineering techniques and most searches were zero-shot. This could mean that users become very reliant on the particular model they use and form an opinion bubble shaped by the LLM. While for a search engine the same cannot be excluded, the user is at least presented with sources from various sides at once. After all, search engines exist for long enough that the technique of searching on them is widely known.

Yet, contrasting studies also exist. [23] found that users of ChatGPT™ were both faster and more accurate compared to Google™ users. However, Google™ users' results were better when they visited more sources - for ChatGPT™ source count was found to be insignificant. Additionally, [50] finds in their second experiment that when provided with a confidence

[1] search engine results page

level for the LLM's answer, users' responses increase in accuracy dramatically while the users remain satisfied with the system. [19] have already tried to create a framework for the quality of human-LLM interaction. The framework distinguishes engagement from the actual information retrieval. In their demonstration, the researchers found that users can be nudged to constructively engage with LLMs instead of becoming reliant. These findings may suggest that LLMs are simply underexplored in their use scenarios. And, more importantly, they suggest that context may matter. The difference between a search engine and an LLM is larger than it seems and tacit knowledge from years of using SERP may not translate directly into chatbots.

2.3. Query Formulation and Prompt Engineering

By this point, people are aware of how to type search queries on platforms like Google™ - it should be brief and consist of keywords. With LLM chatbots, the story is different. Prompt engineering refers to the term behind optimizing queries to receive better answers from an LLM [11]. These may range from role-playing to asking for a slow thinking mode, and are dependent on the AI model used. There is not a single correct solution.

[36] finds that users understand the basics of prompt engineering but are not aware of the details. Most participants used descriptions instead of instructions, few used a modifier and many did not specify the style for an image generation task. This holding back of information may restrict the context of the LLM. In traditional search engines, the user may click several links, use filters on websites, etc. With an LLM, however, users must be upfront, since all context must be inferred from their query. [28] evaluated what a good prompt is. Among many, they found manner, clarity, task decomposition and a logical structure to be important factors. Yet, none of these have to be considered for a simple web search. Furthermore, [16] outline that not all prompts are of the same type. They list four modes - text-based prompting, UI-interactions, context-based prompting, task-delegation. For this study, context-based interaction is of particular interest.

Some automated solutions were proposed in several works to reformulate queries automatically before passing them onto the LLM [58, 33, 56, 57, 30]. The works present (LLM-based) algorithms which refine the query and make it more understandable for the LLM. [18] propose a tool to help identify systematic LLM output differences. Together, these tools may aid users to optimize their queries in the long-term. Yet, providing sufficient context to the LLM is still a burden that the user must learn to accept. An interesting partial solution is presented by [1], who proposes automatically including user knowledge history and conversational history into the LLMs current context.

2.4. Human Information Seeking

This research is based on the patterns of human information gathering, specifically during web searches and LLM usage. In [39], Information Foraging Theory (IFT) is conveniently presented and adapted for LLM chatbots. IFT, with roots in psychology, assumes that humans seek information with systems comparable to (and based on) their former food-foraging systems.

In IFT, humans treat information as prey. This ”prey” is organized in patches [39]. In simple words, people seek to find things which are scattered across multiple sources. On the internet, the easiest example of different sources is a set of different websites, but it can also be databases or documents. People treat the information by two metrics - how useful it is (=value) and how difficult it is to get (=cost). Then, people try to find as much useful information with as little effort as possible. This is a maximizing strategy.

Humans can either search within one source or they could search between multiple sources [39]. The longer people search a single source, the less new useful information they gain - at some point there is nothing left to read on one website. When people decide to search multiple sources, they use cues (=scent) to understand what source may have the next valuable piece of information.

Yet not all information-based tasks are the same. [3] found that three distinct decision-making tasks depend on different combinations of uncertainty and time horizon. When being uncertain which option among several choices is the best, time and the degree of uncertainty highly influence a person's decisions. When time is plentiful or uncertainty is high, people are more likely to explore with their decisions. The same holds true when people decide how much information they gather prior to making a large decision. With enough time and/or large uncertainty, people research more. When they are sure what option is best or they are running out of time, people stop checking and pick the best current choice. For “foraging” tasks the situation is different. When deciding between which information source to use (i.e. the current one, or a new one), uncertainty has a smaller role. Time, however, remains a large factor. If a person has a lot of time left, they are more likely to explore more sources and vice versa.

[39] specifically hypothesize that LLMs differ from search engines for information seeking. They suspect that:

1. LLMs require more precise input to retrieve information, compared to search engines.
2. LLMs act as a single, evolving source, while search engines require users to navigate multiple sources (e.g., websites, databases).
3. If information is spread across multiple sources, retrieving it via an LLM is more resource-intensive than using a search engine.
4. LLMs limit users to themselves, whereas search engines access a wider range of external sources.
5. Cost-value estimations in LLMs include accuracy and relevance.

Combining these findings and thoughts, it stands to reason that we must test whether people actually search differently with an LLM as opposed to a search engine. Would one option capture more context and provide more relevant answers? Or does the difference in tool not matter for accurate and satisfactory results?

2.5. Maximizers vs. Satisficers

In the 1950s, Herbert A. Simon first introduced the idea of ”satisficers” - people that differ from ”optimizers” [49, 46]. According to Simon, some problems have incomplete context (e.g. incomplete information). For these problems, a ”satisficer” would seek a solution which is solely satisfactory (i.e. a solution that meets certain criteria or better) [46]. This is in stark contrast with an ”optimizer” or ”maximizer”, which would try to maximize a certain pay-off function [49]. The maximizer profits from the best possible result, yet maximizing is often unrealistic when context of the task is not fully known. Simon gave examples ranging from business [47] to chess [48]. Chess specifically is a good example, as players often consider only a handful of possible decisions instead of computing all possible outcomes. They hence find a satisfactory next move instead of finding the outright best one. However, by doing so they save time and mental resources.

In order to measure whether a person is a maximizer or satisficer, a personality matrix can be used. In 2002, Barry Schwartz et al. introduced a 13-item maximization scale which they inferred from an initial 42-item questionnaire [43]. The maximization scale was based on behavioral examples of maximizing and high standards. It also had a complementary 5-item regret scale. The maximization scale created in [43] consisted of three distinct maximization dimensions: *high standards* (the desire to accept only the absolute best outcome), *alternative search* (the behavioral tendency to exhaustively explore all available options), and *decision difficulty* (the frustration or anxiety experienced when forced to choose). These terms, however, were not yet properly named in [43] - they were only later formally defined by [35].

Using this framework, Schwartz et al. found that maximizers more often engaged in social comparison than satisficers, both when evaluating potential decisions and achieved outcomes [43]. This occurs because the concept of the “best” option is amorphous - it cannot be measured objectively. More generally, regret, perfectionism, and depression were found to strongly correlate with the maximizing tendency, whereas happiness, optimism, and satisfaction with life had a negative correlation [43]. In addition, the authors demonstrated that maximizers are generally more regretful and less happy with their consumer decisions. In an ultimatum bargaining game, maximizers were less satisfied with their outcomes than satisficers [43].

In 2016, Cheek and Schwartz provided a breakdown of several revised versions of the scales [7]. The enhanced versions addressed developments and criticisms, for instance by removing the decision difficulty components since decision difficulty was more of an outcome than a cause. Moreover, the new versions addressed the issue that maximizers were unfairly portrayed as less happy/satisfied in the 2002 version [14, 7]. The 2016 paper reasserts that there exist two maximization components – *goal* and *strategy*. Goal refers to Simon’s original idea of making the best choice [49, 46, 7]. The strategy component consists of exhaustive alternative search to achieve said best choice [7]. In simple terms, maximization goal captures how strongly someone wants the best possible outcome rather than merely a good-enough one. Maximization strategy captures how someone tries to reach that outcome: for example, by comparing many alternatives, continuing to search, and delaying a decision until they feel they have explored enough options.

Thus, a person may strongly want the best outcome without necessarily engaging in extensive search behavior, or they may search extensively as a way of avoiding a premature decision.

Research found that people who are satisficers differ from maximizers not only in their decision making, but also in result perception. [25] analyzed that maximizers are often optimists and have a tendency for consistency and risk aversion. Meanwhile, in [43], using the 2002 scale, it is concluded that the maximizers are also more likely to be unhappy about the decisions they make and that they have a tendency to regret their decisions. While this result is now outdated, [6] and [53] have (using newer scales) found that maximizing as a *strategy* still correlates with negative experiences and depression respectively. [20] confirm this on a practical job search example. Graduating students that had maximizing tendencies received jobs which paid 20% more than the ones found by their satisficer peers. However, the same maximizer students were less happy about their job search results.

It is thus not far-fetched to assume that personality may play a role when users are interacting with different recommender systems. Scientific literature presents mixed findings. [9] found that when faced with a choice based on reviews, maximizers constructed their decisions differently from satisficers and may be more prone to suffer from "choice overload". More generally, choice overload has also been observed in recommender system settings in [5], where larger sets of attractive recommendations increased choice difficulty without necessarily improving satisfaction with the chosen item. Yet, in [34] the role of choice overload is minimized based on an e-commerce example. The researchers found no correlation with choice set size and choice satisfaction, though they acknowledged that in general, maximizers are less satisfied with their final choice. Meanwhile, [22] found no difference between maximizers and satisficers when it came to decision-making behavior while using recommender systems. The story is similar with AI. [45] found that maximizers are more likely to deeply integrate AI into their workflow (and use it overall), assuming there is a high degree of perceived usefulness. [24], on the contrary, finds that maximizers exhibit a greater reactance (=aversion) towards AI recommendations. The author suggests that this is linked to a reduced feeling of autonomy and obstruction of their main search.

For our research, it is important to keep these differences in mind. We may find out that one personality group prefers classic searching to LLMs because of control and greater parallel-search capabilities while the other group may lean towards the simplicity and conversational style of an LLM recommendation.

3. The Present Study

The general goal of this MSc thesis is *to compare how LLM-based search and traditional search engines support information retrieval.* Building on the distinction between deeper interaction with a single evolving source and broader navigation across multiple sources, the study focuses on two outcome measures: result accuracy and user satisfaction. Hence, the following two research questions arise:

1. Is there a difference in result accuracy between using an LLM and a search engine for research tasks?
2. Is there a difference in user satisfaction between using an LLM and a search engine for research tasks?

As reviewed above (section 2), we can see that LLMs and search engines present data in different ways: LLMs may act as a single source of truth, whereas search engines provide various results in a list, for instance. Hence, LLMs may motivate deeper search, whereas search engines promote a broader search, impacting what results users achieve. Moreover, the differences in the user interface (UI) and user experience (UX) between the two platforms may be more or less pleasant for particular users, especially when combined with the fact that the style of research is different for each person. Hence, we create two hypotheses:

H1 *Using an LLM for information retrieval yields **worse** result accuracy than using a search engine.*
H2 *Using an LLM for information retrieval yields **better** user satisfaction than using a search engine.*

Figures 1 and 2 showcase these hypothetical main effects.

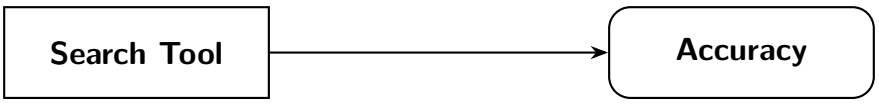


Figure 1: A diagram showing that *Search Tool* may affect *Accuracy*.

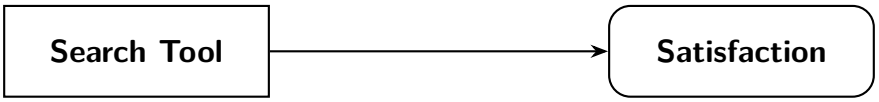


Figure 2: A diagram showing that *Search Tool* may affect *Satisfaction*.

Based on the revised personality matrix from [7], we can assume that maximizers and satisficers can be defined by two components - goal and strategy. Researching information for a certain task incorporates both these components. People have a goal (fulfilling the task by finding an answer) and a strategy (how they search). Hence, the following two research questions arise:

3. Does the personality type have an effect on result accuracy?
4. Does the personality type have an effect on user satisfaction?

From these research questions, four additional hypotheses can be created:

H3a *Maximization goal* ***moderates*** *the relationship between search tool and result accuracy.*
H3b *Maximization strategy* ***moderates*** *the relationship between search tool and result accuracy.*
H4a *Maximization goal* ***moderates*** *the relationship between search tool and user satisfaction.*
H4b *Maximization strategy* ***moderates*** *the relationship between search tool and user satisfaction.*

Figures 3 and 4 showcase how goal and strategy may moderate the main effect between the search tool and accuracy/satisfaction. Importantly, goal and strategy are two separate moderation variables as this reflects how personalities are classified in [7]. Similar to [6] and [53], we may find that only one of these constructs moderates the main effect.

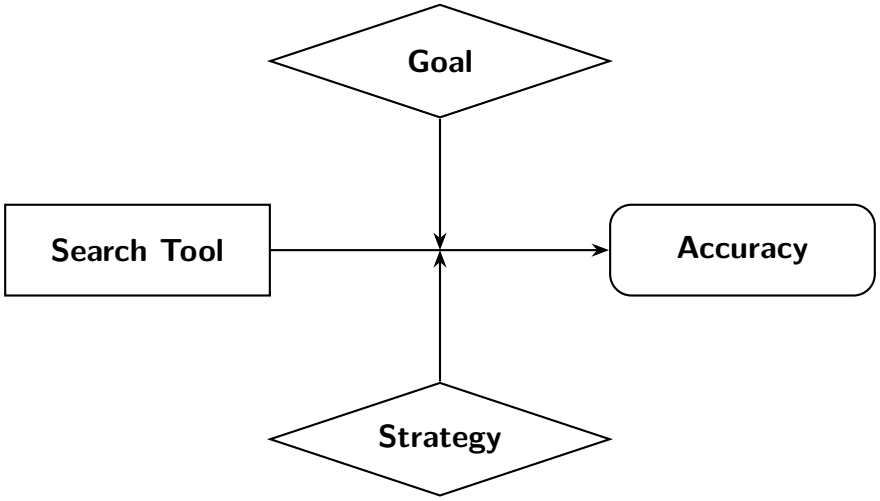


Figure 3: A diagram showing *Goal* and *Strategy* moderating the relation between *Search Tool* and *Accuracy*.

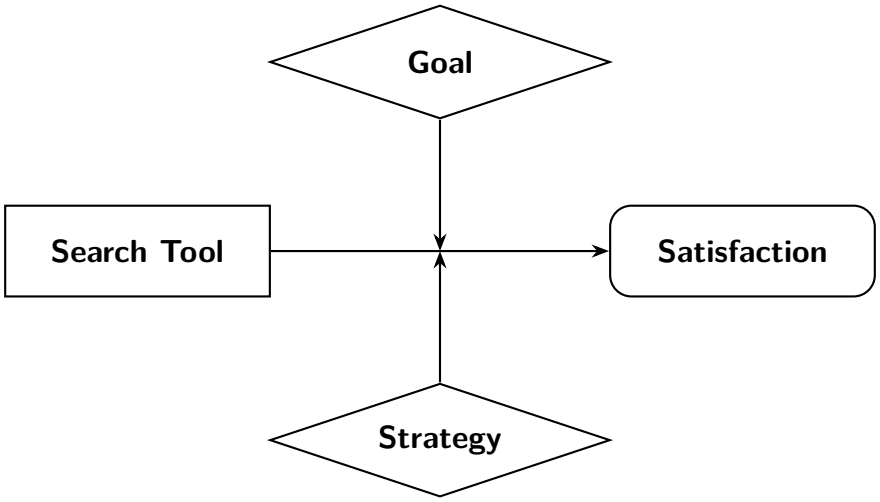


Figure 4: A diagram showing *Goal* and *Strategy* moderating the relation between *Search Tool* and *Satisfaction*.

4. Search Tasks

This section shows the search tasks which participants of the experiment have to complete. In this study participants complete all tasks sequentially (one-by-one). Effectively, each task consists of a person researching a given topic and answering a questionnaire (survey) about it. The detailed explanation of the experimental setup (e.g. what participants are given, where they write their answers, etc.) can be found in the sections 5 and 6.

4.1. Task 1: Person Search

This search task consists of identifying a person using the limited information provided about them. After the person is identified, the participants must look into their academic work to answer this task's questions. The exact task looks like this:

> Task A student from TPM at the TU Delft has gotten the "The Steven Hoogendijk Best Thesis Award 2019-2020 from the Batavian Society for Experimental Philosophy" award. This student has written an interesting thesis.
> You are allowed to use your search mechanism, visit any websites the search mechanism provides, use the "image" and "video" sections of the search mechanism. You are not allowed to go to websites directly (bypassing the search mechanism). You are not allowed to use any search mechanisms other than the one provided to you (e.g. you cannot use ChatGPT).
>
> Q1 What is the title of the student's master thesis?
>
> A1 Language Inquiry for Personalized Mental Health Chatbots.
>
> Q2 One of the professors who supervised this graduate later published a paper based on the graduate's work. Where was that paper published?
>
> A2 Frontiers in Digital Health.

From anecdotal experience, any search tool should perform well for people and academic search tasks. The difficulty from this task stems from two main components, however. Firstly, the "Steven Hoogendijk" award stems from a niche group. Hence, the resources about this award are sometimes limited. In addition, we chose a person who received the award for the 2019-2020 academic year. In this year, the COVID-19 pandemic caused some delays: the student was awarded their prize in 2021, simultaneously with the laureate of 2021. In other words, two students received their awards in the same year. Lastly, the student lists this prize on their LinkedIn™, as shown in figure 5. However, there is a typo in the name ("sociaty") which complicates finding this extract via online search.
All in all, this task examines participants' text search skills within a limited information field.

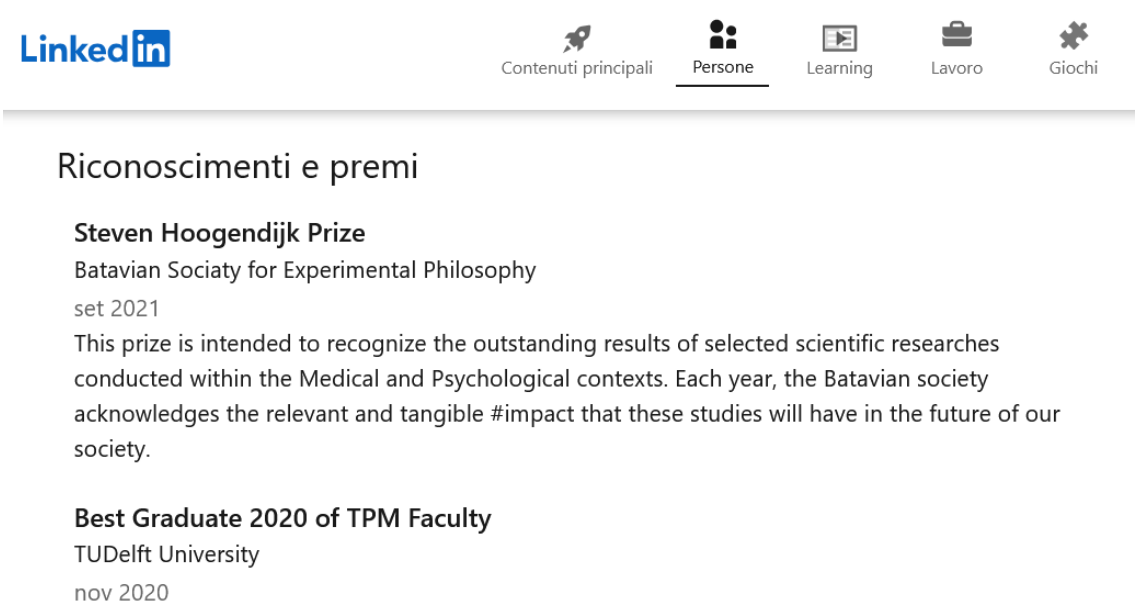
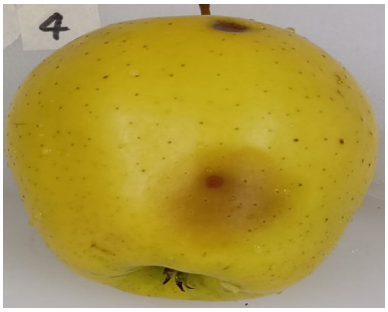


Figure 5: Screenshot of the LinkedIn™ profile of a former TPM student who won the Steven Hoogendijk Prize in 2021.

4.2. Task 2: Apple Disease Identification

The "Apple Disease Task" consists of the participant receiving several images of apples with various conditions. Figure 6 depicts example conditions. Images 6a and 6b show an apple with alternaria rot, whereas figure 6c shows an apple with bull's eye rot.

(a) Example of alternaria rot.



(b) Second example of alternaria rot.

(c) Example of bull's eye rot.

Figure 6: Images of apples with various conditions.

Participants receive the following questions for the apples and are expected to provide the following responses:

Task Provided to you are several images of apples. The images can be found on your desktop, in the ”Apple Images” folder. The apples have certain conditions, for instance a disease. Your task is to identify each apple’s condition with the help of your search mechanism.
You are allowed to use your search mechanism, visit any websites the search mechanism provides, use the “image” and “video” sections of the search mechanism. You are not allowed to go to websites directly (bypassing the search mechanism). You are not allowed to use any search mechanisms other than the one provided to you (e.g. you cannot use ChatGPT).

Q What is the condition of the apples in the first two images? You only have to write down the condition itself. You do not have to provide an explanation.

A Alternaria rot.

Q What is the condition of the apple in the third image? You only have to write down the condition itself. You do not have to provide an explanation.

A Bull’s eye rot.

This task differentiates itself by making the participant examine images instead of text. Participants can thus use reverse image search with the LLM to make it identify the apple’s condition or they can manually search for common conditions with the search engine and compare results. Another layer of complexity is added by the fact that apple conditions (such as diseases) often look similar to an untrained eye. Hence, there exists a high risk of misidentifying an infection as a physiological disorder. This risk exists not just for the user - LLMs are also at risk of incorrectly analyzing the image.

5. Instrument Design

In this section, the instrument design of the experiment is shown. In our case, "instrument" refers mainly to the search tool that participants use. In addition, the tool for surveying the participants (i.e. collecting the answers they provide for the tasks) is also presented. For detailed information about research methods, such as participant selection and experiment design, see section 6.

For this study, participants are given one of two search tools – DuckDuckGo™ or OpenRouter™. Participants tackle two unrelated tasks from section 4 using this particular tool, enabling us to observe consistency or shifts in their strategies.

5.1. DuckDuckGo™

DuckDuckGo™ is a privacy-centric search engine similar to Google™. It operates as an alternative search engine to the better-known competitors with a focus on privacy. Despite this, DuckDuckGo™ is very customizable. This customization makes it a great choice for this study. DuckDuckGo™ allows us to turn off AI-answers and AI-nudges so that the participants can only use the classic search engine environment. Nonetheless, DuckDuckGo™ has the same widgets people expect from Google™ – maps, weather, etc. An example screenshot of DuckDuckGo™ is shown in figure 7. Here, a familiar search interface can be found. The search query is at the top. Answers are presented in a list right below the query. An instant answer (called "Search Assist") can be found before the hyperlinks to other websites. In addition to web search, participants have the ability to use the "image" and "video" search features of DuckDuckGo™, along with any other options the search engine provides. In addition, DuckDuckGo™ allows users to filter websites based on region (e.g. search Dutch sites exclusively), tweak safe search level and exclude older sources from search results. It is safe to assume that anybody who is tech literate will have no issue using this familiar user interface.

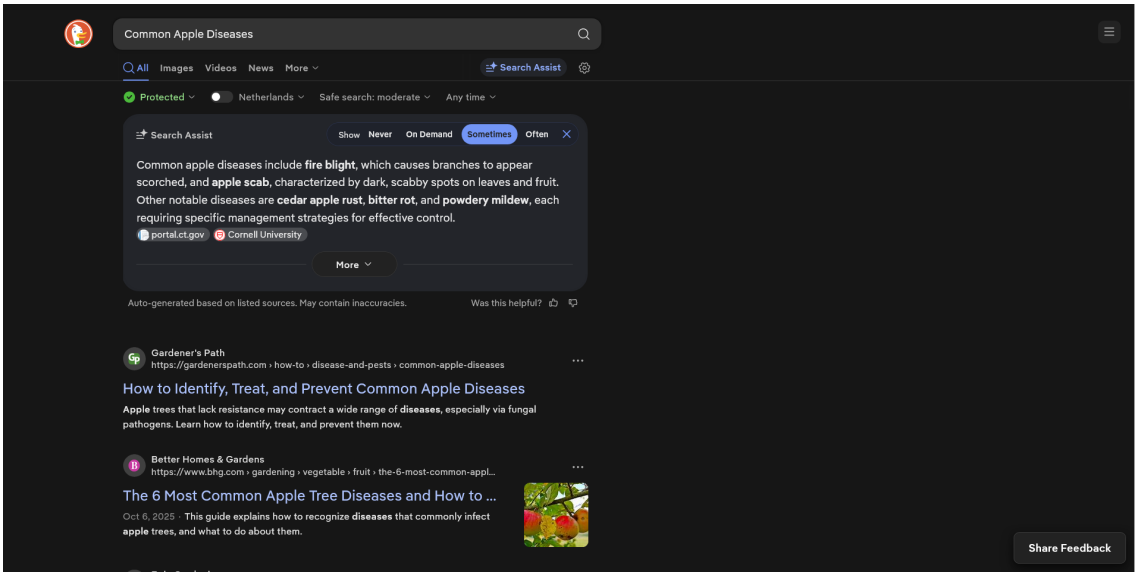


Figure 7: Screenshot of the DuckDuckGo™ search engine. Search results for common apple diseases are shown.

For our use case, DuckDuckGo™ is set to mainly default settings (incl. standard safe search, no regional filters, etc.). Notably, we deactivate any AI-buttons. This is to avoid users accidentally going to Duck.ai™ (an LLM chatbot) if they were meant to only use DuckDuckGo™. "Search Assist" is also deactivated, as it is an AI-based summary feature. We explicitly disallow participants to change any settings during the experiment. However, participants may use the aforementioned filters (e.g. video search, regional search). Notably, using these filters is not required – it may simply be convenient for some participants.

5.2. OpenRouter™

OpenRouter™ describes itself as "The Unified Interface For LLMs"[2]. It is primarily a service that allows people to use a single API to call various high-end models from different providers. Recently, however, OpenRouter™ launched a "chat" section on its website. This chat section acts as an LLM frontend. By frontend, we mean that the company does not expect us to use a model of its own. Instead, OpenRouter™ chat acts as a convenient middleman. Through it, we get access to models from OpenAI™, Anthropic™, Mistral AI™, etc., all in one convenient chat window. Crucially, because these models are accessed through API calls, we deal with increased rate limits. This allows for multiple people to use one account simultaneously without seeing warning messages or errors. Moreover, OpenRouter™ has a failsafe option - if one LLM provider blocks us, OpenRouter™ switches providers automatically, continuing a user's chat.

Figure 8 depicts a screenshot of the website. At the top, a previously entered query can be found. The query asks about the disease of an apple. A picture of said apple is attached to the query. Below, an LLM-generated answer is provided. The answer lists a likely diagnosis. We also see the "Mistral Medium 3" tag before the answer meaning that Mistral's Medium 3 model was used for the answer. If an answer was generated with the usage of web search, relevant links are provided below the response. At the bottom of the screen, there is a text field for typing the next query. Typing the query will continue the already existing chat. There are also buttons next to the text field for tuning the AI's chat memory, attaching files from the computer, and selecting web search. The left side-window of the interface allows the user to start a new chat, and change the LLM model. Overall, the interface should look familiar to anybody who has used LLMs in the past. The only quirks that have to be explained are explicit enabling of web search and model choice.

[2] https://openrouter.ai/

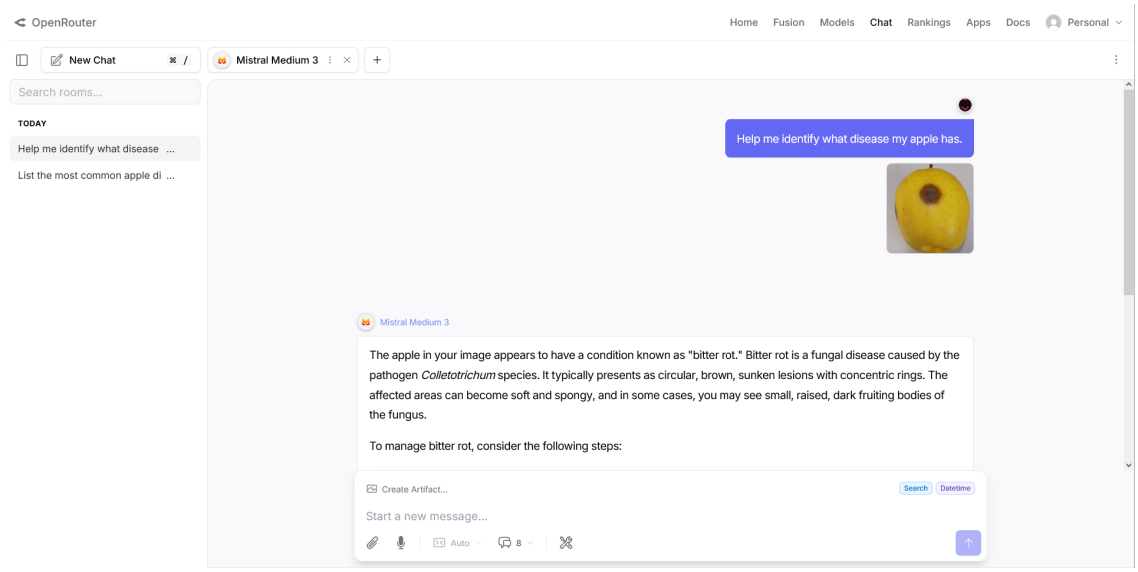


Figure 8: Screenshot of the OpenRouter™ LLM chat. The Mistral Medium 3 model is asked to identify an apple's disease.

For the purposes of this research, participants are given the default version of OpenRouter™ – no settings are changed. Participants are explicitly told not to change any settings on their own as this might affect the outcome of the study in unexpected ways. OpenRouter™ allows users to select which LLM they want to use – there is a choice between models from OpenAI™, Mistral AI™, Meta™, etc. For consistency purposes, we have selected the same model for all users, that being the Grok™ 4.3 model from xAI™. Initially, we intended to use a more widely adopted model, such as Claude™. However, during testing, several models available through OpenRouter™ did not work reliably when web search was enabled. Grok™ 4.3 was selected because it handled web search consistently and was one of the newer models with reasoning capabilities available at the time of the experiment. To parse users' chats, we simply used the "export chat" feature of OpenRouter™. This downloads the chat to our computer in a JSON format. Notably, once the chats are deleted on OpenRouter™ or once browser cookies are wiped, these chats are permanently gone. Hence, participants are told not to delete their chats.

5.3. Microsoft Edge™

Computer browsers require little introduction nowadays. For the purposes of this study, the Edge™ (=MS Edge™ or simply Edge™) browser from Microsoft™ is used. The browser is based on Chromium™, the same browser engine that the dominant Google™ Chrome™ browser uses. This assures us that websites load and behave like they are supposed to. Edge™'s defining feature is its ability to export browser history without needing to install any extensions. The history exports to a CSV file which allows us to easily parse it with any script. This CSV export is critical for the study, as it allows for the precise, quantitative extraction of browser navigation patterns, such as page visits, required to analyze participant behavior. Lastly, MS Edge™ also has a "browser profiles" feature. This feature allows us to split sessions for each participant on the same computer, if necessary. While

we ended up not needing it, researchers with a tight amount of computers can isolate history data for each participant without risking mixing up the data.

Figure 9 shows the default setup of MS Edge™ (tabs and URL bar on top, website at the bottom). Figure 10 provides a screenshot of the history tab. For this study, MS Edge™ is run on default settings. Participants are not allowed to change settings during the experiment to keep results consistent.

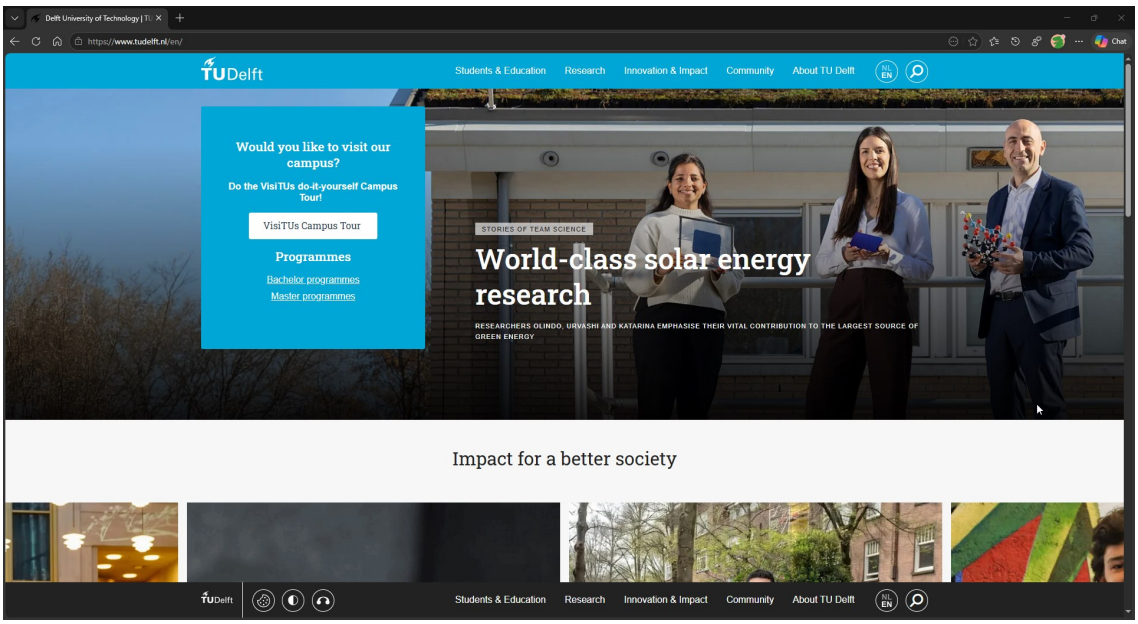


Figure 9: Screenshot of the MS Edge™ browser with a TU Delft tab open.

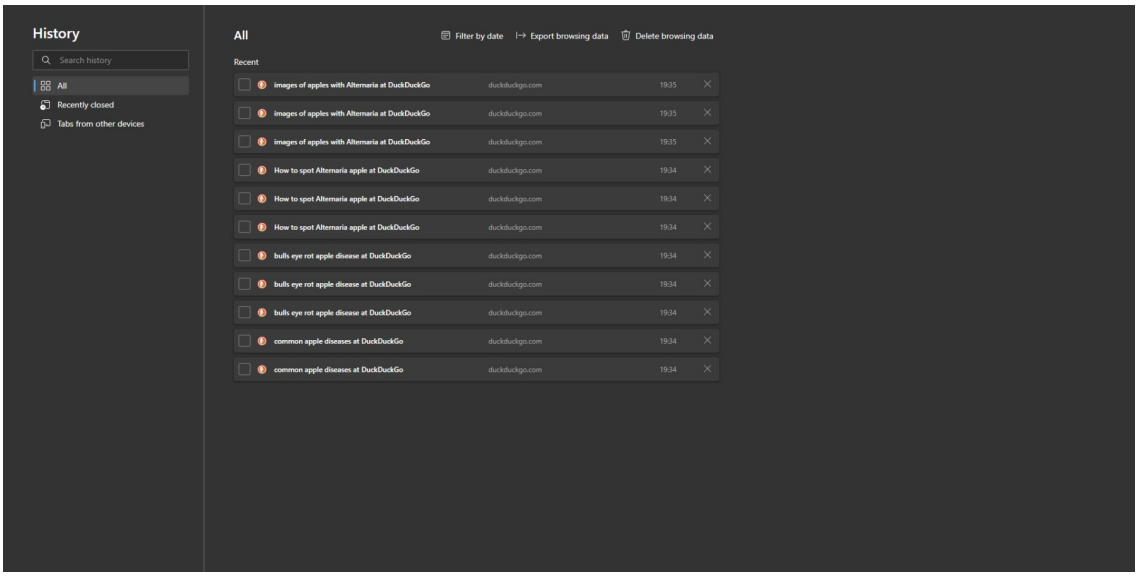


Figure 10: Screenshot of the history page of the MS Edge™ browser. The history can be exported to a CSV file.

5.4. MS Forms™

In this study, participants are expected to fill out several surveys. For this purpose, we use Microsoft Forms™ (=MS Forms™, Forms™). MS Forms™ is a good fit for two reasons - it is simple to use and it is part of TU Delft's Microsoft™ license, meaning that it has extra privacy protections compared to competing products.

MS Forms™ can be opened within the Edge™ browser. Its user interface is intuitive with linearly laid out questions and clear answer options. An example of a personality survey using MS Forms™ can be seen in figure 11.

Participants may only submit their answers to a survey once. All answers (except follow-ups) are flagged as "required" meaning participants cannot skip a question.

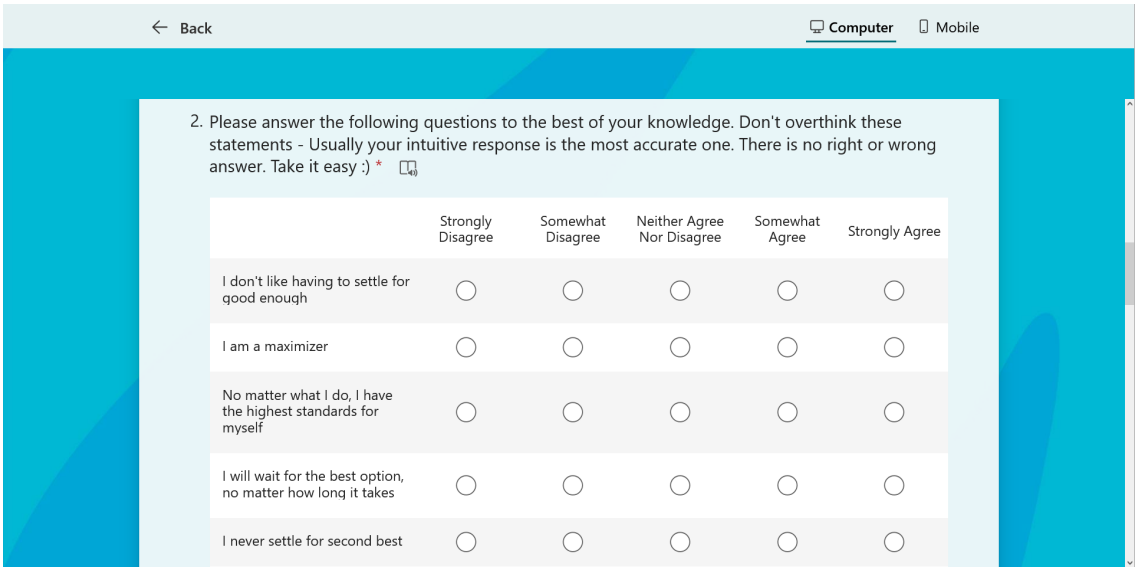


Figure 11: Screenshot of Microsoft™ Forms™ survey about a person's personality.

5.5. Process Flow

After outlining the most important instruments used in the study, this subsection explains the general flow designed for the participants. Figure 12 illustrates this flow for better understanding.

At the start of the session, participants receive a short introduction to the study, read the information sheet, and provide consent to participate. They are then informed of the search tool which they have been assigned, either DuckDuckGo™ or OpenRouter™. The participants are then seated at a computer in the TU Delft computer room. On the computer, an MS Edge™ window is prepared in advance for use during the experiment.

Participants begin by navigating to the pre-survey in MS Forms™ from the favorites bar. After completing and submitting this survey, they close the window and receive the first search task, which is provided on paper for convenience. They reopen Edge™ and navigate to the assigned search tool, which participants use to complete the task. Once they have finished the task, or once the time limit has expired, they go to the post-task survey in

MS Forms™ by clicking the shortcut in the favorites bar. In the survey, they submit their answer and complete the post-task questions, after which they close the window. This sequence is repeated for each of the two search tasks. After the final task and post-task demographics survey have been completed, the session ends.

During the session, additional software runs in the background to support data collection. Inputlog records user interactions such as keystrokes and mouse actions, while vokoscreenNG records the participant's screen as a backup measure in case a browser window is closed prematurely. Participants are informed of these recordings before the experiment begins, but they do not interact with either program directly.
Inputlog data is meant for a future study. For the purpose of this study, the data was not used.

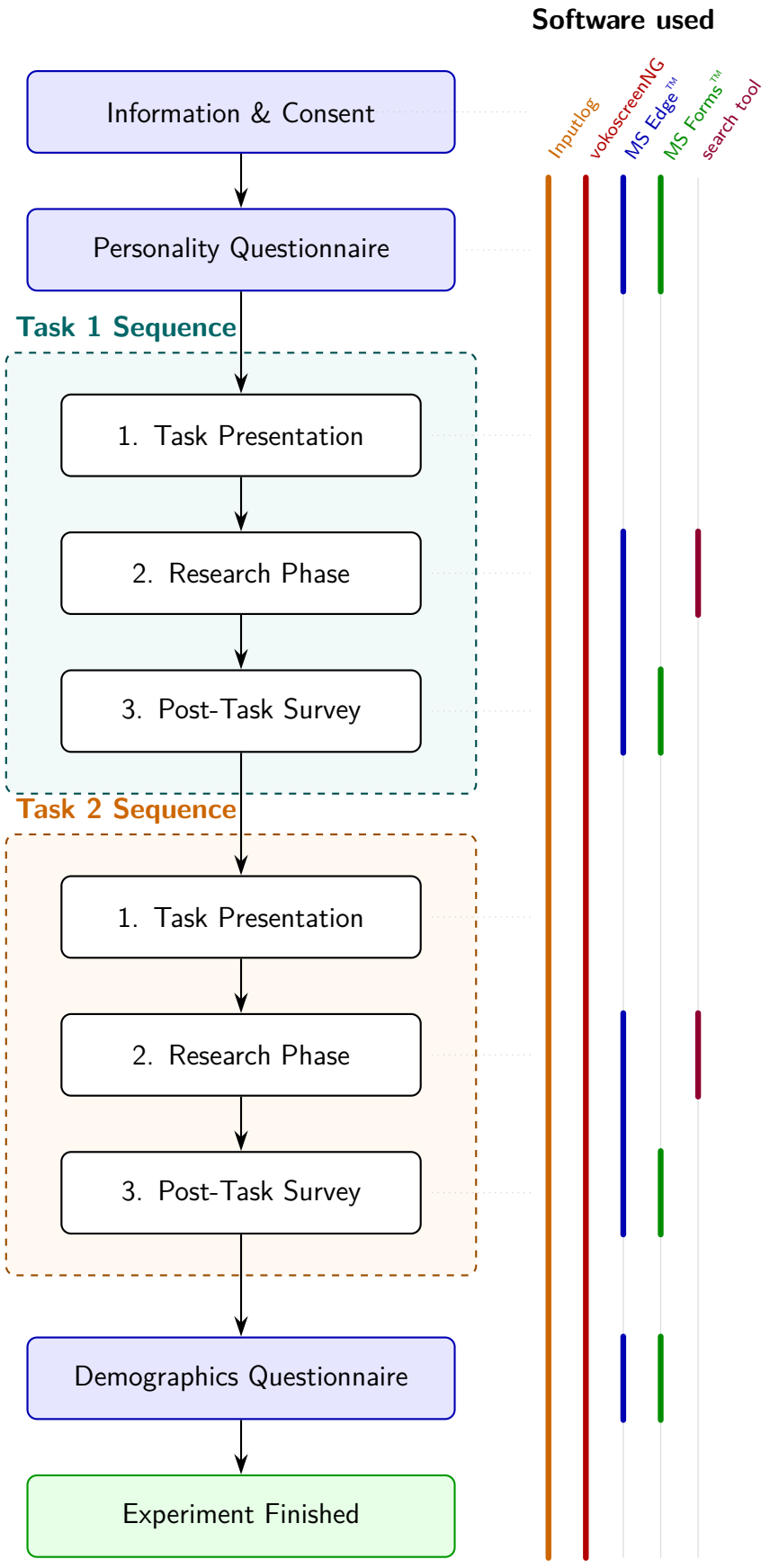


Figure 12: Sequential flow of the experimental procedure with parallel software-use lines.

6. Research Method

This section outlines the methodological design of the study. First, the experimental design is described, followed by the experimental procedure. Next, the manipulation is presented. Finally, the measures of the experiment are explained.

6.1. Ethics Approval

This experiment-based study and its protocol received official approval by the Human Research and Ethics Committee (HREC) of the TU Delft [case number 6537]. All collected data is handled and stored in accordance with the project's approved Data Management Plan.

6.2. Experimental design and participants

The study employed a mixed experimental design with one between-subjects manipulation and a repeated-measures component. The between-subjects manipulation was the search tool: participants were randomly assigned to use either DuckDuckGo™ or OpenRouter™ throughout the experiment. The repeated-measures component consisted of two search tasks completed by each participant using the same assigned tool. In addition, individual differences in maximization were measured before the tasks using the scales that [7] considered to be the best fits for goal and strategy. For goal, this is the MTS-7 [12] scale. For strategy, the MI[3] [52] was used.

The study was primarily conducted in a computer room at TU Delft[4]. After receiving information about the study and providing informed consent, participants accessed two web applications in Microsoft Edge™: MS Forms™ and their assigned search tool. Participants were randomly assigned to either DuckDuckGo™ (search engine) or OpenRouter™ (LLM) before beginning the experiment. They first completed the maximization questionnaires in MS Forms™, after which they performed the two search tasks using their assigned tool. Each task had a maximum time limit of 10 minutes, excluding filling out the related survey. Following each task, they completed a post-task survey in MS Forms™ in which they submitted their answer and rated their satisfaction with the search process.

Participants were recruited at campuses of the TU Delft and Leiden University. Predominantly, the subjects were undergraduate students from both universities. The experiment was conducted in-person in a designated computer room on the university campus[4]. The experiment was conducted on an exclusively voluntary basis with participants having the option to opt-out at any time (including in the middle of the experiment). Participants were offered 15 euros to participate in the experiment. The demographic composition of our sample can be seen below in tables 1 to 6.

[3] The alternative search scale specifically was used.

[4] As an additional recruitment measure, a mobile approach was also used where participants were approached with a set up laptop on campus. The rest of the process remained the same.

Table 1: Search tool distribution.

Search Tool	Frequency
LLM chat (OpenRouter™)	14
Search Engine (DuckDuckGo™)	14

Table 2: Gender distribution.

Gender	Frequency
Male	20
Female	8

Table 3: Age distribution.

Age range	Frequency
18–25	23
26–30	5

Table 4: Education distribution.

Education	Frequency
High School	12
Bachelor's Degree or Equivalent	10
Master's Degree or Equivalent	6

Table 5: Search engine use frequency.

Search engine use	Frequency
Daily	24
A few times a week	3
Once or twice a week	1

Table 6: AI use frequency.

AI use	Frequency
Daily	18
A few times a week	5
Once or twice a week	4
Once or twice a month	1

6.3. Procedure

Upon arrival, participants were welcomed, introduced to the general context of the study regarding information retrieval, and asked to sign an informed consent form. Following consent, participants were seated at a computer where they first completed a pre-study survey hosted on MS Forms™. This pre-survey administered the MTS-7 and MI personality scales to measure participants' maximization strategies and goals. Once the pre-survey was submitted, participants were seamlessly directed to their randomly assigned search tool.

During the core phase of the experiment, participants completed the search tasks using their assigned environment. The technical execution and on-screen process flow for these tasks are detailed in section 5.5. Participants navigated between the task prompts and the search tool independently, with the system silently logging their queries, search behavior, and time metrics to capture their behavioral navigation patterns.

Immediately after concluding each individual search task, participants were required to complete a post-task survey on MS Forms™. This final survey asked participants to input their definitive answer to the task (used to measure accuracy) and included Likert-scale questions to assess their subjective satisfaction with the search process. Once all tasks and post-surveys were completed, participants were thanked for their time and dismissed.

6.4. Manipulation

The independent variable in this experiment is the type of search tool used by participants. This variable was manipulated by randomly assigning each participant to one of the two options - DuckDuckGo™ or OpenRouter™. The participant used the same search tool across all tasks. This setup allowed us to compare whether the type of search tool affects task outcomes, in this case accuracy and satisfaction. In addition, we could measure whether the maximizing traits (goal and strategy) influence these effects.

As a manipulation check, participants were asked in each survey which search tool had been assigned to them. This question was used to verify whether participants were aware of the experimental condition they were in (DuckDuckGo™ or OpenRouter™). Responses to this item allowed us to identify whether participants correctly recognized their assigned search tool.

6.5. Measures

6.5.1. Dependent measures

The dependent measures are result accuracy and user satisfaction. Result accuracy was chosen because it should effectively represent a search tool's usefulness - a search tool is only useful when you find what you were searching for. User satisfaction was chosen as a measure because it should represent whether a person is willing to actually make use of the tool - a powerful search tool nobody uses can be considered a failure. Combined, these two measures give us a holistic view of a search tool's utility.

Result accuracy was operationalized as the correctness of the participant's submitted answer for each search task. For each task, responses were coded according to a predefined answer key, after which an overall accuracy score was computed by counting the amount of correct responses for each participant.

User satisfaction was operationalized as participants' self-reported evaluation of the search experience after each task. Specifically, users filled in a modified version of the "User satisfaction" scale from [26] (based on [8]). Participants did not receive an answer key or any feedback about the correctness of their answers before rating their satisfaction. Responses were recorded using the 7-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". The scale can be found in the appendix, in table A.1. The scale showed a Cronbach's $\alpha = 0.860$ for task 1 and $\alpha = 0.949$ for task 2, meaning it is reliable in both cases.

6.5.2. Maximizing Goals

Maximizing goals refer to the tendency to strive for the best possible outcome when making decisions, as per [49, 46, 7]. In this study, maximizing goals were measured using the MTS-7 scale, originally developed by [12] and outlined by [7] as a good fit. Example items include "No matter what I do, I have the highest standards for myself." and "I never settle.". The scale can be found in the appendix, in table A.2. Responses were recorded using the 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". The maximizing goal scale showed a Cronbach's $\alpha = 0.645$, which falls below the commonly accepted threshold of 0.70. Inspection of the component loadings suggests that some

goal items may contribute less strongly to the underlying construct. In particular, *goal_4* and *goal_6* showed relatively low loadings on component 2, indicating that these items share less variance with the rest of the scale. This may partly explain the lower reliability of the goal scale. Nonetheless, all questions related to maximizing goals fit into the same component, as can be seen in table 7.

6.5.3. Maximizing Strategies

Maximizing strategies refer to the tendency to engage in extensive comparison and search behavior during decision making, as per [7]. In this study, maximizing strategies were measured using the alternative search scale, based on [52]. Example items include "I take time to read the whole menu when dining out." and "I take the time to consider all alternatives before making a decision.". The scale can be found in the appendix, in table A.3. Responses were recorded using the 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". The maximizing strategy scale showed a Cronbach's $\alpha = 0.862$, meaning it is reliable. When conducting principal component analysis (seen in table 7), all questions related to maximizing strategy fit into the same component.

Table 7: Component Loadings for personality questionnaire.

	Component 1	Component 2	Uniqueness
goal_1	0.143	**0.657**	0.548
goal_2	−0.082	**0.735**	0.453
goal_3	−0.139	**0.680**	0.517
goal_4	0.194	**0.252**	0.899
goal_5	0.074	**0.665**	0.553
goal_6	0.008	**0.330**	0.891
goal_7	0.156	**0.481**	0.745
strat_1	**0.497**	0.249	0.691
strat_2	**0.716**	−0.005	0.488
strat_3	**0.744**	−0.249	0.384
strat_4	**0.790**	−0.013	0.375
strat_5	**0.516**	0.014	0.733
strat_6	**0.687**	−0.381	0.382
strat_7	**0.670**	0.070	0.547
strat_8	**0.559**	−0.089	0.680
strat_9	**0.610**	0.301	0.538
strat_10	**0.581**	0.297	0.575
strat_11	**0.713**	0.145	0.471
strat_12	**0.521**	0.093	0.720

7. Results

7.1. Manipulation Check

To verify whether the manipulation of search mechanism was successful, participants' self-reported assigned search mechanism was compared with the experimental assignment key. Participants reported which tool they used in every questionnaire, totaling four responses per participant. The table comparing responses against the assignment key can be found in the appendix (table B.1).

All reported responses corresponded to the assigned search mechanism. Participants assigned to the search engine condition reported using the search engine, while participants assigned to the LLM condition reported using the LLM. Therefore, the manipulation check showed perfect recall accuracy across participants and measurement points.

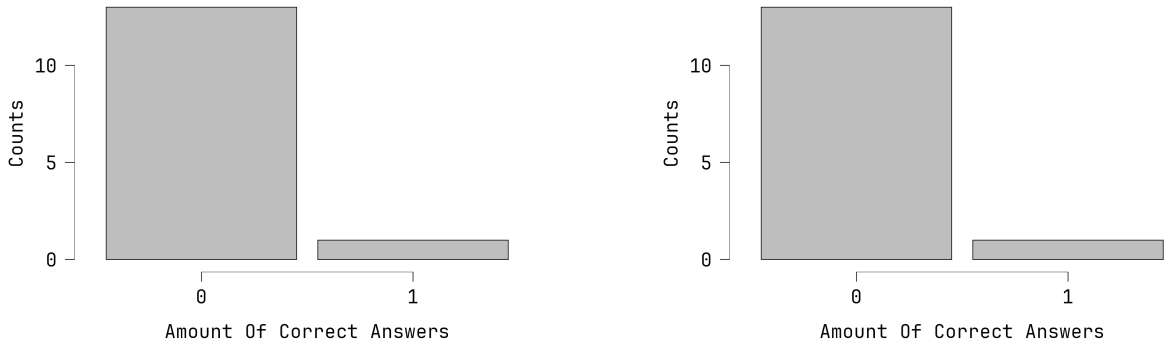
As there were no incorrect responses, the manipulation check outcome contained no variance. For this reason, no inferential model was estimated. The manipulation check therefore confirms descriptively that participants correctly recognized their assigned search mechanism.

7.2. Descriptive Statistics

Before proceeding with testing the hypotheses, it is important to look at the basic details around the experiment. Starting with the outcomes of each task, figure 13 shows the number of correct responses for task 1 while figure 14 shows the same for task 2. For task 1, only one person from either cohort (Search Engine vs. LLM) managed to provide one correct answer. For task 2, no one using the search engine could provide a correct answer while the LLM users successfully provided one correct answer three times. Notably, not a single participant from any group managed to provide more than one correct answer throughout the whole experiment. Table 8 shows the information in tabular form.

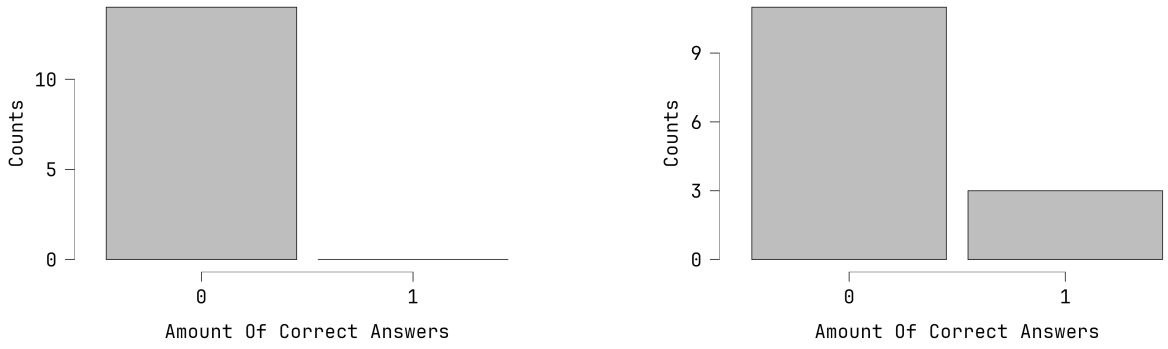
Table 8: Descriptive statistics for accuracy.

	Accuracy task 1		Accuracy task 2	
	DuckDuckGo™	LLM	DuckDuckGo™	LLM
0 correct	13	13	14	11
1 correct	1	1	0	3
2 correct	0	0	0	0
Total correct answers	1	1	0	3



(a) DuckDuckGo™. **(b)** OpenRouter™.

Figure 13: Amount of correct answers for task 1.



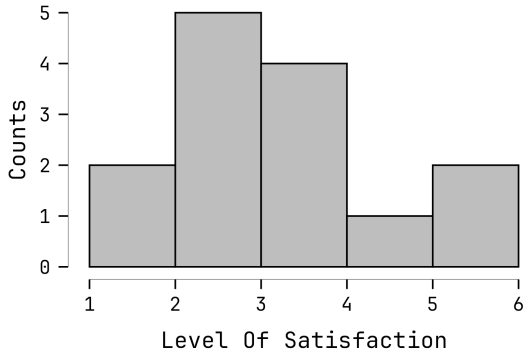
(a) DuckDuckGo™. **(b)** OpenRouter™.

Figure 14: Amount of correct answers for task 2.

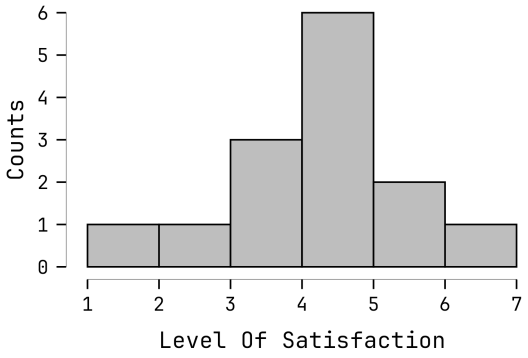
In terms of satisfaction, there is a larger difference between the groups. For both tasks, shown in figures 15 and 16, we observe that the LLM users rated their satisfaction as substantially higher than the search engine users. This is especially noticeable in task 2, where the mean for LLM users was 5.143 while DuckDuckGo™ users had a mean of just 3.339. For task 1, the difference is smaller, but still measurable - the mean of LLM users was 4.304 while DuckDuckGo™ users rated their satisfaction 3.429 on average. Table 9 shows the information in tabular form.

Table 9: Descriptive statistics for satisfaction.

	Satisfaction task 1		Satisfaction task 2	
	DuckDuckGo™	LLM	DuckDuckGo™	LLM
Mean	3.429	4.304	3.339	5.143
Std. Deviation	1.238	1.305	1.625	1.343
Minimum	1.250	1.750	1.000	1.750
Maximum	5.500	7.000	5.500	6.250

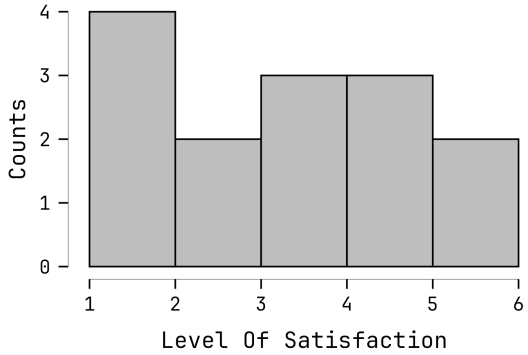


(a) DuckDuckGo™.

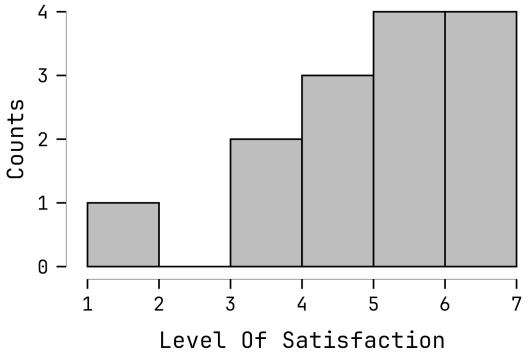


(b) OpenRouter™.

Figure 15: Satisfaction distribution for task 1.



(a) DuckDuckGo™.



(b) OpenRouter™.

Figure 16: Satisfaction distribution for task 2.

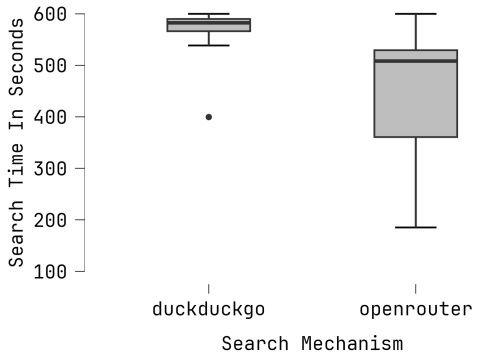
In addition to the main outcome variables, exploratory behavioral indicators were collected to describe how participants searched. These indicators allow us to indirectly assess the breadth and depth of participants' search behavior. In this thesis, breadth of search is operationalized as the number of external page visits and the number of unique external domains visited during a task. These measures indicate how widely participants moved across information sources. Depth of search was less directly captured, but query count is used as an exploratory indicator of iterative engagement with the assigned search tool. We consider the search tool to be the starting point of the search process. Therefore, if visiting multiple websites indicates a broader search across sources, sending multiple queries to the assigned search tool can be interpreted as deeper engagement with the same topic. Search time is reported separately as a general indicator of task duration, rather than as a direct measure of breadth or depth. Because these behavioral measures were not part of the main hypotheses, they are interpreted descriptively rather than inferentially.

Figure 17 shows the descriptive statistics for task 1 completion. On average, LLM users required 442 seconds to complete task 1, whereas search engine users needed nearly the maximum amount of time at 566 seconds. Participants using DuckDuckGo™ also created nearly twice as many queries - 7.6 on average compared to the mean of 4.4 queries for

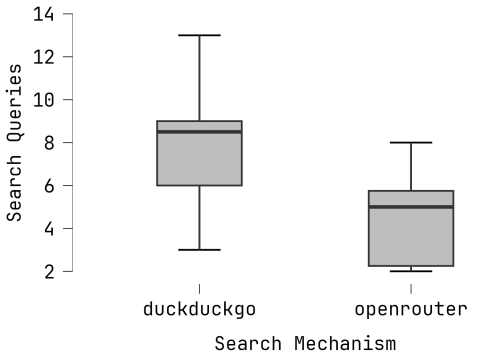
OpenRouter™ users. In terms of external websites accessed, the average search engine user accessed 23.3 websites, 6 times more than the mean of 3.9 for AI users. ”External websites” is a flaky metric, as some webpages can create multiple history entries through bot-checks, URLs for page sections, and other means. Hence, we also compute the amount of unique domains participants visited. This should provide a more realistic image of the users’ online navigation. In terms of unique domains, the difference remains similar. 8.5 websites were accessed on average by search engine users, compared to a mere 1.6 websites by participants using an LLM. Detailed statistics can be seen in table 10.

Table 10: Descriptive statistics for task 1.

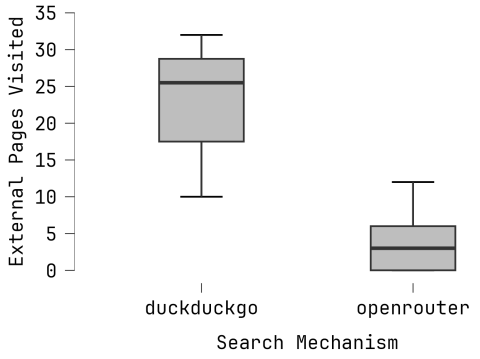
Variable	Search Mechanism	Mean	SD	Min.	Max.
Search time (s)	DuckDuckGo™	566.1	51.45	399.6	600.0
Search time (s)	OpenRouter™	442.0	143.6	185.2	600.0
Query count	DuckDuckGo™	7.57	2.82	3	13
Query count	OpenRouter™	4.36	2.02	2	8
External page visits	DuckDuckGo™	23.29	7.05	10.00	32.00
External page visits	OpenRouter™	3.86	3.92	0.00	12.00
Unique external domains	DuckDuckGo™	8.50	2.80	5.00	15.00
Unique external domains	OpenRouter™	1.64	1.87	0.00	6.00



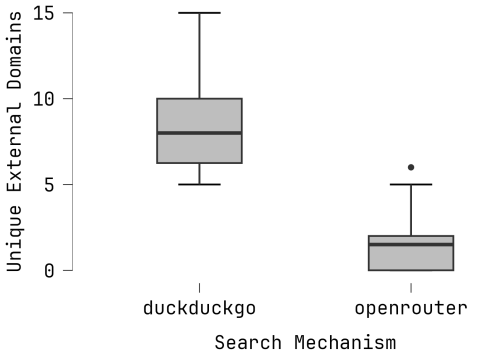
(a) Completion time for task 1.



(b) Search queries for task 1.



(c) External websites visited for task 1.



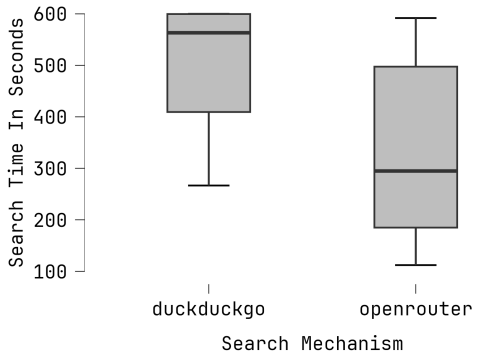
(d) External domains visited for task 1.

Figure 17: Descriptive statistics for task 1.

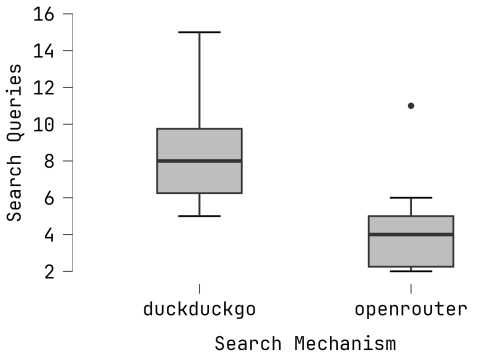
For task 2, similar statistics were created and can be seen in figure 18. Similar patterns to task 1 statistics can once again be found. LLM users finished their task earlier, requiring an average of 337 seconds, compared to search engine users who needed 505 seconds. Roughly twice as many queries were typed by DuckDuckGo™ users - 8.3 as opposed to 4.1 queries by the OpenRouter™ users. The OpenRouter™ group visited only 3.5 websites on average while the DuckDuckGo™ group visited 16.4. Even when only counting unique domain visits, DuckDuckGo™ users take the lead with a mean of 8.6 unique domains compared to OpenRouter™ users' 1.6. Detailed statistics can be seen in table 11.

Table 11: Descriptive statistics for task 2.

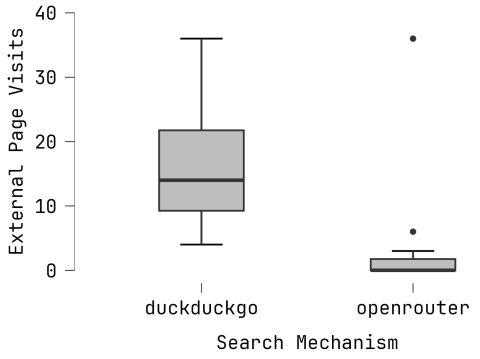
Variable	Search Mechanism	Mean	SD	Min.	Max.
Search time (s)	DuckDuckGo™	504.6	115.2	266.6	600.0
Search time (s)	OpenRouter™	337.3	178.5	112.0	591.9
Query count	DuckDuckGo™	8.29	2.61	5.00	15.00
Query count	OpenRouter™	4.14	2.38	2.00	11.00
External page visits	DuckDuckGo™	16.43	9.48	4.00	36.00
External page visits	OpenRouter™	3.50	9.51	0.00	36.00
Unique external domains	DuckDuckGo™	8.64	3.57	2.00	16.00
Unique external domains	OpenRouter™	1.64	3.18	0.00	11.00



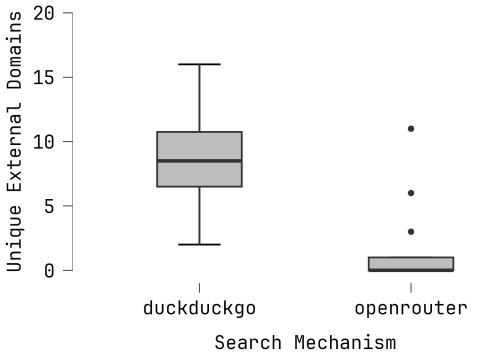
(a) Completion time for task 2.



(b) Search queries for task 2.



(c) External websites visited for task 2.



(d) External domains visited for task 2.

Figure 18: Descriptive statistics for task 2.

7.3. Qualitative Observations of Search Behavior

In addition to the descriptive statistics reported above, the search logs were manually inspected to better understand how participants interacted with their assigned search mechanism. This analysis was exploratory and was not used to test the hypotheses directly. Instead, it was used to provide additional context for the quantitative differences in query count, website visits, unique domains, and task completion time.

Starting with the LLM condition, participants often used conversational queries rather than keyword-based search terms. Instead of entering short phrases, they frequently asked the system full questions or gave it instructions. For example, in task 1, one participant asked: "*A student from the TPM faculty at TU Delft has gotten the [...] Steven Hoogendijk best thesis award 2019-2020 [...] i need you to find this thesis for me*". The same participant later refined the request by asking whether the award existed, whether "*Bataafsch Genootschap der Proefondervindelijke Wijsbegeerte*" referred to the Batavian Society for Experimental Philosophy, and whether the result was a master thesis rather than a PhD thesis. This shows that LLM users often treated the search as an ongoing dialogue, where each new prompt built on the previous response.

This conversational structure was also visible in task 2. LLM users did not only ask for a label, but often asked the system to justify, reconsider, or compare its answer. A representative example is a participant who first asked "*what is the condition of this apple*", then visited one website, and then returned to the LLM with follow-up questions such as "*why do you think it is bitter rot*", "*please double check*", and "*so both of the apples have the same condition?*". This sequence shows that the LLM was used as a diagnostic partner rather than only as a search box. The participant did not simply collect independent webpages, but asked the model to explain its reasoning and to compare the apple images.

A second pattern in the LLM condition was that some participants explicitly tried to control the quality of the generated answer. Several prompts contained phrases such as "*are you sure*", "*make no mistakes*", "*do not hallucinate*", "*provide sources*", or "*give me a more credible source*". This suggests that at least some users were aware of the risk that the LLM could provide an incorrect answer. However, this skepticism was not always accompanied by external verification. Several LLM users continued to interact mainly within the chatbot environment and did not visit external websites. This fits the descriptive statistics, where OpenRouter™ users visited fewer external websites and fewer unique domains than DuckDuckGo™ users across both tasks.

A third pattern was the use of role-based prompting. Some participants attempted to improve the LLM's response by assigning it an expert identity. This was especially common in the apple disease task, where users wrote prompts such as "*You are an expert in botanics*", "*Act like a biologist*", or "*You are apple expert*". These prompts suggest that participants sometimes expected the LLM to perform better if it was instructed to adopt a specialized role. In the context of this study, this is relevant because it shows that LLM users did not only search for information, but also experimented with the form and framing of their prompts.

The DuckDuckGo™ condition showed a different pattern. Search engine users generally re-

lied on shorter keyword-based queries instead of conversational prompts. Their search process was often characterized by incremental keyword reformulation, where participants repeatedly changed or added one or two terms. For example, in task 1, a representative participant began with "*TU Delft steven hoogendijk best thesis award 2019-2020 batavian society for experimental philosophy*", then searched for "*Norbert Kalb*", "*Norbert Kalb tpm tu delft*", "*Norbert kalb thesis TU Delft*", and finally "*Diamond-based quantum networks with multi-qubit nodes*". This sequence shows how search engine users often followed a breadcrumb-like path: they collected names, institutions, titles, and other clues from search results, then reused these clues as new queries.

This broader navigation pattern becomes visible in the representative task 1 chronology. The participant moved between `qutech.nl`, `research.tudelft.nl`, `repository.tudelft.nl`, `quantaneo.com`, `link.aps.org`, `arxiv.org`, `scholar.google.com`, and `catalogue.leidenuniv.nl`. Rather than receiving one synthesized answer, the participant had to reconstruct the answer by moving across institutional, academic, and repository-based sources. This illustrates how DuckDuckGo™ users often followed information across multiple external sources.

For task 2, DuckDuckGo™ users had to approach the apple condition task differently from LLM users. Since the search engine did not directly identify the apple images for them, participants searched for general categories, disease names, and visual examples. One representative participant searched for "*fruit going bad stages wikipedia*", "*stages of ripening apple*", "*stages of molding apple*", and "*stages of decay apple*", while visiting websites such as `thishealthytable.com`, `canr.msu.edu`, `extension.psu.edu`, and `decompositiontime.com`. This shows that search engine users often had to build their own classification framework by comparing the task images to descriptions or examples found online.

The full representative chronologies for both tasks, including the exact logged action values for the LLM and search engine conditions, are provided in appendix C. Here, "representative" means that one participant was selected for each task and condition as an illustrative example of common search patterns observed in the logs, rather than as a statistically representative case.

7.4. Testing Hypotheses

This section reports the statistical tests of the main hypotheses. First, the effect of search mechanism and the moderating role of the two MTB[5] constructs were examined for accuracy using a binomial generalized linear mixed model (count/proportion outcome, repeated tasks). Second, the same predictors were examined for satisfaction using a repeated-measures ANOVA (continuous rating, repeated measurements), with task as the repeated-measures factor. Where significant interactions were found, follow-up analyses were conducted to clarify the pattern of effects.

[5] In the statistical output, MTB_Goal and MTB_Strategy refer to composite scores for the two measured maximization constructs: maximization goal and maximization strategy.

7.4.1. Accuracy

Table 12: Generalized linear mixed model results for accuracy.

Effect	df	χ^2	p
Search_Mechanism	1	0.251	.617
MTB_Goal_Centered	1	0.072	.789
MTB_Strat_Centered	1	0.241	.624
Search_Mechanism * MTB_Goal_Centered	1	1.444	.229
Search_Mechanism * MTB_Strat_Centered	1	0.137	.711

A binomial generalized linear mixed model was conducted to examine the effect of search mechanism and the two personality constructs on accuracy. Task was included as a random effects grouping factor to account for task-specific variability in accuracy, since accuracy was measured separately for each task. The model showed no significant main effect of search mechanism on accuracy, $\chi^2(1) = 0.251$, p = .617. Accordingly, hypothesis H1 was not supported, as LLM use did not result in significantly lower accuracy than search engine use.

The two personality constructs also did not significantly predict accuracy. The main effect of MTB_Goal was not significant, $\chi^2(1) = 0.072$, p = .789, and the same was true for MTB_Strategy, $\chi^2(1) = 0.241$, p = .624. The search mechanism × MTB_Goal interaction was not significant, $\chi^2(1) = 1.444$, p = .229, meaning that hypothesis H3a was not supported. The search mechanism × MTB_Strategy interaction was also not significant, $\chi^2(1) = 0.137$, p = .711, meaning that hypothesis H3b was also not supported. Overall, the results provide no evidence that maximization goal or maximization strategy moderated the relationship between search mechanism and result accuracy.

7.4.2. Satisfaction

Table 13: Between-subjects effects from repeated-measures ANOVA predicting satisfaction.

Cases	Sum of Squares	df	Mean Square	F	p	η_p^2
Search_Mechanism	42.145	1	42.145	18.507	$< .001$	0.457
MTB_Goal_Centered	7.107	1	7.107	3.121	.091	0.124
MTB_Strat_Centered	13.504	1	13.504	5.930	.023	0.212
Search_Mechanism * MTB_Goal_Centered	4.840	1	4.840	2.126	.159	0.088
Search_Mechanism * MTB_Strat_Centered	11.590	1	11.590	5.090	.034	0.188
Residuals	50.098	22	2.277			

A repeated-measures ANOVA was conducted to examine the effect of search mechanism and the two MTB constructs on satisfaction. Task was included as a within-subjects factor, with Satisfaction_1 and Satisfaction_2 representing satisfaction after task 1 and task 2 respectively. Search_Mechanism was included as a between-subjects factor, while MTB_Goal and MTB_Strategy were included as covariates.

The between-subjects effects showed a significant main effect of search mechanism on satisfaction, F(1, 22) = 18.507, p < .001, η_p^2 = .457. Descriptive statistics indicated that satisfaction was higher in the LLM condition than in the DuckDuckGo™ condition for both tasks. Specifically, for task 1, mean satisfaction was 4.304 for LLM and 3.429 for DuckDuckGo™. For task 2, mean satisfaction was 5.143 for LLM and 3.339 for DuckDuckGo™. Therefore, hypothesis H2 was supported: participants using the LLM reported significantly higher satisfaction than participants using DuckDuckGo™.

MTB_Strategy also significantly predicted satisfaction, F(1, 22) = 5.930, p = .023, η_p^2 = .212. Importantly, the search mechanism × MTB_Strategy interaction was significant, F(1, 22) = 5.090, p = .034, η_p^2 = .188. This indicates that the effect of search mechanism on satisfaction depended on participants' level of maximization strategy. Therefore, hypothesis H4b was supported.

Regarding MTB_Goal, the main effect on satisfaction was not significant, F(1, 22) = 3.121, p = .091, η_p^2 = .124. The search mechanism × MTB_Goal interaction was also not significant, F(1, 22) = 2.126, p = .159, η_p^2 = .088. Thus, hypothesis H4a was not supported, indicating that maximization goal did not significantly moderate the relationship between search mechanism and user satisfaction.

Table 14: Simple effects of MTB_Strategy within each search mechanism.

Search Mechanism	Predictor	Sum of Squares	df	Mean Square	F	p
DuckDuckGo™	MTB_Strat_Centered	9.855	1	9.855	3.941	.059
OpenRouter™	MTB_Strat_Centered	0.115	1	0.115	0.046	.832

To further examine the search mechanism × MTB_Strategy interaction, a follow-up simple slopes analysis was conducted within the repeated-measures ANOVA framework using dummy-coded search mechanism variables. First, the simple slope of MTB_Strategy was examined within each search mechanism. When DuckDuckGo™ was the reference condition, the effect of MTB_Strategy approached significance, F(1, 24) = 3.941, p = .059. When LLM was the reference condition, the effect of MTB_Strategy was not significant, F(1, 24) = 0.046, p = .832. This suggests that maximization strategy was more relevant for satisfaction among DuckDuckGo™ users than among LLM users, although the within-DuckDuckGo™ association only approached conventional significance.

Table 15: Effect of LLM versus DuckDuckGo™ at different levels of MTB_Strategy.

MTB_Strat_Centered Level	Predictor	Sum of Squares	df	Mean Square	F	p
Low (-1 SD)	LLM vs DuckDuckGo™	26.727	1	26.727	10.689	.003
Avg	LLM vs DuckDuckGo™	32.959	1	32.959	13.182	.001
High (+1 SD)	LLM vs DuckDuckGo™	5.541	1	5.541	2.216	.150

Next, the effect of search mechanism was tested at low and high levels of MTB_Strategy. For this analysis, search mechanism was dummy coded as LLM_Code, with DuckDuckGo™

coded as 0 and LLM coded as 1. The effect of search mechanism was significant at low strategy, $F(1, 24) = 10.689$, $p = .003$. At high strategy, the effect of search mechanism was no longer significant, $F(1, 24) = 2.216$, $p = .150$. Together, these results suggest that the LLM condition produced higher satisfaction than DuckDuckGo™ among participants with low MTB_Strategy, but that this difference disappeared among participants with high MTB_Strategy.

As a parallel check, the same analyses were repeated with the 0/1 coding reversed, so that DuckDuckGo™ was coded as the target comparison instead of LLM. Because this coding reverses the direction of the same comparison, the F-values and p-values are identical to the LLM-coded analysis reported in Table 15.

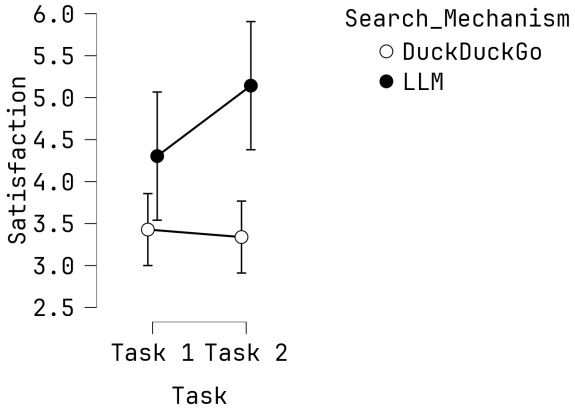


Figure 19: Descriptive plot of satisfaction across task 1 and task 2 by search mechanism. Error bars represent 95% confidence intervals.

Figure 19 descriptively shows that participants in the LLM condition reported higher satisfaction than participants in the DuckDuckGo™ condition across both tasks. The difference appears larger in task 2 than in task 1.

8. Discussion

8.1. Scientific Relevance

This study contributes to the literature on LLM-based information retrieval by comparing traditional search engines and LLM chatbots not only in terms of task accuracy, but also in terms of user satisfaction and individual differences in maximization. This comparison was conducted across two distinct task contexts: a text-based person-search task and an image-based disease-identification task. In doing so, the study connects research on human information seeking, Context-Aware Recommender Systems, and maximization theory.

First, hypothesis H1 predicted that using an LLM would lead to lower result accuracy than using a search engine. This hypothesis was not supported: the study did not find a significant difference in accuracy between the LLM and search-engine conditions. This finding does not clearly replicate earlier work suggesting that LLM-based search can reduce accuracy when users rely on erroneous generated answers [50]. At the same time, the result should not be interpreted as evidence that LLMs and search engines are equally accurate in general. Accuracy was very low across both experimental tasks, which suggests that task difficulty limited the ability to detect differences between conditions. This connects to information foraging theory, where search outcomes depend strongly on the distribution of information, the uncertainty of the task, and the effort required to move between information sources [3, 39]. The result may therefore suggest that accuracy differences between LLMs and search engines may be highly task-dependent rather than purely tool-dependent.

Second, hypothesis H2 predicted that using an LLM would lead to higher user satisfaction than using a search engine. This hypothesis was supported. Participants in the LLM condition reported higher satisfaction than participants in the DuckDuckGo™ condition. This finding is consistent with earlier studies showing that users often evaluate LLM-based or conversational search systems positively, even when accuracy risks remain present [50, 27]. The result also supports the broader argument from recommender systems research that objective performance and user experience should be treated as separate evaluation dimensions [21, 59]. In this study, the LLM did not significantly improve accuracy, but it did improve satisfaction. This reinforces the idea that the perceived usefulness or pleasantness of an information retrieval system cannot be inferred from correctness alone.

Third, hypotheses H3a and H3b predicted that maximization goal and maximization strategy would moderate the relationship between search mechanism and result accuracy. These hypotheses were not supported. Neither maximization goal nor maximization strategy significantly changed the effect of search mechanism on accuracy. This finding contributes to the mixed literature on maximizers and recommender systems. Some studies suggest that maximizers may search differently, experience choice overload, or respond differently to recommendations [9, 24, 45]. Other work, however, finds limited differences between maximizers and satisficers in recommender system decision behavior [22]. The present findings are closer to the latter interpretation for objective task performance. However, because the accuracy outcome showed little variation, the current

study cannot rule out the possibility that maximization affects accuracy in tasks with a broader range of difficulty.

Fourth, hypotheses H4a and H4b examined whether maximization goal and maximization strategy moderated the relationship between search mechanism and satisfaction. Here, the findings differed between the two maximization constructs. Maximization goal did not significantly moderate satisfaction, whereas maximization strategy did. This distinction is theoretically relevant because in [7] it is argued that maximization should not be treated as a single construct, but as a combination of separate components such as goal and strategy. The present study supports this distinction: merely wanting the best outcome did not explain differences in satisfaction across tools, but the tendency to search extensively did. This effect is parallel to [6] and [53], which associated maximization strategy with more negative decision-related experiences, an outcome closely related to satisfaction. More specifically, the satisfaction advantage of the LLM was found among participants low in maximization strategy, but disappeared among participants high in maximization strategy. This pattern fits the idea that high-strategy maximizers may place greater value on extensive comparison and visible alternatives, which are more naturally supported by traditional search engines than by LLMs acting as a single evolving source [39]. Therefore, the LLM interface may be less uniquely satisfying for users who prefer extensive alternative search.

Lastly, the descriptive behavioral measures also help connect the findings back to the distinction between breadth and depth of information retrieval introduced earlier. Participants using DuckDuckGo™ appeared to show broader search behavior than participants using OpenRouter™, as they visited more external websites and unique domains across both tasks. In contrast, OpenRouter™ users stayed primarily within the LLM environment. This pattern is consistent with the idea that traditional search engines encourage between-source navigation, whereas LLMs concentrate the search process within a single conversational interface [39]. Depth was less directly captured, but query count may be interpreted as an exploratory indicator of iterative engagement with the search tool. Here too, DuckDuckGo™ users generated roughly twice as many queries as OpenRouter™ users across both tasks. Finally, users of OpenRouter™ completed their tasks more quickly than DuckDuckGo™ users. Importantly, the behavioral measures were used descriptively rather than as primary dependent variables. They should therefore be interpreted as exploratory indicators of search behavior, rather than direct evidence that one tool produces deeper or broader information retrieval.

The qualitative inspection of the representative search chronologies provides additional support for this interpretation. In line with information foraging theory, DuckDuckGo™ users appeared to move between multiple information patches, using search results, website visits, names, domains, and article titles as cues for where to search next [39]. This was visible in the representative person search chronology (table C.1), where the search engine user moved from the award name to Norbert Kalb, TU Delft pages, repository pages, and academic publication sources. OpenRouter™ users, in contrast, more often remained within one evolving conversational source. Their prompts were usually phrased as full questions or commands, and later prompts often built on previous responses. This

fits the idea that LLM-based search changes the information seeking process from navigating across sources to interacting with a single system that summarizes and reformulates information for the user [39]. At the same time, this pattern also connects to earlier concerns about over-reliance and verification in LLM-based search [50, 27]. Some LLM users explicitly asked the system to "*double check*", "*make no mistakes*", or provide sources, but this was not always accompanied by independent website verification. Therefore, the chronologies help explain why OpenRouter™ may have felt more satisfying while not producing a clear accuracy advantage: the system reduced the amount of manual navigation, but did not necessarily improve the reliability of the final answer.

Overall, the scientific contribution of this study lies in showing that the comparison between LLMs and search engines should not be reduced to accuracy alone. The results support earlier work showing that LLM-based search can be experienced as more satisfying, while also suggesting that this satisfaction advantage is conditional on users' decision-making style. In particular, maximization strategy appears to be more relevant than maximization goal when explaining satisfaction with different search mechanisms. This extends existing research by connecting LLM-based information retrieval with maximization theory and by showing that individual differences may shape how users experience new search technologies.

8.2. Practical Relevance

This study focuses on a rapidly developing area of information retrieval. As LLM-based tools become increasingly integrated into search, workplace productivity, and recommender systems, understanding how users experience these tools compared with traditional search engines has practical relevance for organizations. The present findings are especially relevant for three areas: the evaluation of LLM tools, the design of Context-Aware Recommender Systems, and the development of verification-oriented search interfaces.

First, the results suggest that LLMs can improve user experience, but should not be evaluated on satisfaction alone. Participants in the LLM condition reported higher satisfaction than participants using DuckDuckGo™, even though the LLM did not significantly improve accuracy. For organizations, this means that integrating LLM-based tools into knowledge work may improve the perceived quality of the work experience. This fits with applied evidence from [13], who found that consultants using AI completed tasks 25.1% more quickly and completed 12.2% more tasks, when these tasks were within the AI system's capabilities. However, the same paper warns about AI becoming less accurate when the task was outside of its frontier. In practical terms, organizations should therefore avoid treating user satisfaction as sufficient evidence that an LLM tool is reliable. Instead, LLM deployments should be evaluated through both subjective measures, such as satisfaction and usability, and objective measures, such as correctness, source quality, and downstream error rates.

Second, the findings are relevant for Context-Aware Recommender Systems and online service platforms such as e-commerce. Participants low in maximization strategy showed

a satisfaction advantage in the LLM condition, whereas this advantage disappeared among participants high in maximization strategy. This suggests that recommender systems may benefit from adapting not only to product context or browsing history, but also to users' decision-making style. Users who prefer direct guidance may benefit from an LLM-based assistant that summarizes options or provides a recommendation. In contrast, users high in maximization strategy may prefer interfaces that support comparison, filtering, and visible alternatives. E-commerce is a useful example: a platform that combines conversational recommendation with traditional comparison tools may serve multiple user types at once. This connects to applied work on personalization, where McKinsey™ reports that faster-growing companies generate substantially more value from personalization than slower-growing competitors [32]. Similarly, AI-enabled "next best experience" systems have been argued to improve customer satisfaction by 15-20% when they deliver the right interaction at the right moment [29]. The present study adds that such personalization may also need to account for how users prefer to search and decide.

Third, the findings suggest that LLM-based search tools should be designed with verification and control mechanisms. In this study, the LLM condition produced higher satisfaction, but not significantly higher accuracy. This combination is practically important because users may experience LLM search as pleasant or efficient while still producing incorrect or incomplete answers. Prior work already shows that LLM-based search can create a mismatch between satisfaction and accuracy or verification [50, 27]. Recent applied work on professional AI use adds a further concern: even when professionals actively try to validate LLM outputs, the system may respond with increasingly persuasive explanations rather than neutral correction [40]. For organizations, this means that LLM tools should not only optimize for convenience or satisfaction. They should also make verification structurally easy by providing source links, uncertainty indicators, alternative results, or prompts that encourage users to check generated answers outside of the conversational loop.

Taken together, these findings also provide a qualified answer to the argument introduced at the start of this thesis that LLMs may improve Context-Aware Recommender Systems through richer human-computer interaction [42]. The higher satisfaction in the LLM condition supports this idea from a user experience perspective, but the lack of an accuracy advantage means that the present study cannot conclude that LLM interaction improves objective retrieval quality. Instead, the practical value of LLMs may lie in improving the interaction layer through which users search, express preferences, and provide context, especially for users with lower maximization-strategy scores. Organizations and system designers should therefore explore this direction, provided that LLM-based systems include verification and comparison mechanisms to reduce the risk of erroneous or incomplete recommendations.

8.3. Limitations

As this study functioned as a pilot experiment for a larger study, several methodological limitations should be considered when interpreting the results. These limitations concern the choice of search platforms, the rapidly changing nature of LLM-based systems,

the design of the search tasks, the experimental environment, and the sample composition.

First, the comparison between a search engine and an LLM was necessarily operationalized through two specific platforms: DuckDuckGo™ and OpenRouter™. This choice improved experimental control, because DuckDuckGo™ allowed AI-based search features to be disabled and OpenRouter™ provided API-level access to LLMs with web-search functionality. However, this also means that the findings are not necessarily generalizable to all search engines or all LLM interfaces. Prior research comparing generative search systems and traditional search engines similarly depends on the exact systems being compared, because differences in interface design, source presentation, and model behavior can influence both search behavior and satisfaction [50, 27, 23]. In the present study, this limitation is especially relevant for the search engine condition, because several participants were less familiar with DuckDuckGo™ than with more widely used search engine Google™. In addition, DuckDuckGo™ did not provide reverse image search, which may have disadvantaged search engine users in the image-based apple disease task. Future work should therefore compare multiple search engines and LLM interfaces, or select platforms with more functionally equivalent features.

Second, the LLM condition depended on a specific model and interface configuration. OpenRouter™ was selected because it allowed multiple participants to use an LLM with web search capabilities during group sessions, but the available model and web search functionality changed after piloting. As a result, the final experiment was tied to Grok™ 4.3 rather than to LLMs in general. This is a common challenge in research on generative AI systems, where models and interfaces develop quickly and may differ substantially across time, providers, and deployment settings. Prior work also highlights that LLM-based search outcomes depend on system design choices, such as whether users are provided with confidence information, source links, or interaction support [50, 27, 23]. Therefore, the present findings should be interpreted as evidence about the specific experimental configuration used in this study, rather than as a stable estimate of all LLM-based search systems. Future studies should document model versions, interface settings, and web search behavior in detail, and should ideally replicate the design across multiple LLMs.

Third, the search tasks appear to have been more difficult than intended. Accuracy scores were low across both tasks: only a small number of participants provided any correct answer, and no participant answered more than one question correctly. This floor effect limits the sensitivity of the accuracy analyses, because when very few participants obtain correct answers, it becomes difficult to detect differences between search mechanisms or moderation effects of personality. The difficulty of the tasks may have been useful for observing search behavior under uncertainty, but it reduced the usefulness of accuracy as an outcome measure. This limitation is especially important because information seeking behavior is known to depend on task characteristics such as uncertainty, time horizon, and the distribution of information across sources [3, 39]. Future studies should therefore pilot tasks more extensively and include a wider range of task difficulties. A design with easier, medium, and difficult tasks would make it possible to avoid floor effects while

still examining how users behave when information is harder to retrieve.

8.4. Future Work

This study fulfills the purpose of our pilot research. However, future researchers may want to build on top of this foundation.

A first direction for future work would be to conduct a follow-up study with a broader and more carefully balanced task set. In the present study, accuracy scores were very low, which made it difficult to determine whether search mechanism truly had no effect on accuracy or whether the tasks created a floor effect. A logical study 2 would therefore keep the comparison between LLM-based search and traditional search engines, but include tasks with varying levels of difficulty. This would make it possible to examine whether the relative performance of LLMs and search engines depends on task difficulty, uncertainty, and the distribution of information across sources, as suggested by information foraging theory [3, 39]. In this sense, task difficulty should be understood as one part of task context rather than as context as a whole. Source quality, source availability, and conflicting perspectives may also be relevant dimensions to explore. Ideally, the study would also include a larger sample of participants for greater statistical power.

Second, this study points out that maximizing strategy influences the level of user satisfaction. An obvious follow-up would be seeing what happens when this trait is integrated into a search tool directly. For instance, as suggested in section 8.2, maximizers may receive an open system with many filters, whereas satisficers may be presented with an AI-assistant on the same platform. By splitting the user experience based on maximizing behavior, studies may observe a higher overall satisfaction with a given tool. This would connect the current findings to Context-Aware Recommender Systems research, where the usefulness of context depends on whether it is relevant, representational, and capturable [2], as well as to prior work on maximizers in recommendation settings [9, 22].

8.5. Conclusion

The rise of LLMs has introduced new ways of researching information. As these systems rapidly develop and become increasingly integrated into everyday workflows, it becomes important to understand how they compare with traditional search engines. Previous research has compared LLMs and search engines with varying results. This thesis built on that work by examining whether the search tool affects result accuracy and user satisfaction, and whether users' maximization tendencies moderate these effects.

To investigate this, a mixed experimental design was developed. Participants completed two previously unseen search tasks using either an LLM or a traditional search engine. After each task, they submitted their answers and reported their satisfaction with the search process. In addition, participants completed a personality questionnaire measuring two dimensions of maximization: maximization goal and maximization strategy.

The results showed no significant difference in result accuracy between the LLM and search engine conditions. However, accuracy was low across both tasks, meaning that this finding should be interpreted cautiously. In contrast, the search tool did significantly

affect user satisfaction: participants using the LLM reported higher satisfaction than participants using the search engine.

The study also showed that maximization strategy moderated the satisfaction effect. The LLM satisfaction advantage was present among participants low in maximization strategy, but disappeared among participants high in maximization strategy. Maximization goal did not show the same moderating effect. This suggests that the tendency to search extensively may be more relevant for understanding satisfaction with search tools than the general goal of wanting the best possible outcome.

Overall, this thesis shows that the comparison between LLMs and traditional search engines should not be reduced to accuracy alone. LLMs may offer a more satisfying search experience, but this advantage is not universal and may depend on users' decision-making style. As LLMs become more common in information seeking, future research should further examine when these tools improve the search process, when they create risks, and for which users they are most suitable.

References

[1] Praveen Acharya. “Towards Effective Modeling and Exploitation of Search and User Context in Conversational Information Retrieval”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM ’23. ACM, Oct. 2023, pp. 5161–5164. DOI: 10.1145/3583780.3616005.

[2] Gediminas Adomavicius et al. “Context-aware recommender systems: From foundations to recent developments”. In: *Recommender systems handbook*. Springer, 2021, pp. 211–250.

[3] Bruno B. Averbeck. “Theory of Choice in Bandit, Information Sampling and Foraging Tasks”. In: *PLOS Computational Biology* 11.3 (Mar. 2015). Ed. by Paul Schrater, e1004164. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004164.

[4] Mahta Bakhshizadeh. “Supporting Knowledge Workers through Personal Information Assistance with Context-aware Recommender Systems”. In: *Proceedings of the 18th ACM Conference on Recommender Systems*. RecSys ’24. Bari, Italy: Association for Computing Machinery, 2024, pp. 1296–1301. ISBN: 9798400705052. DOI: 10.1145/3640457.3688010. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/3640457.3688010.

[5] Dirk Bollen et al. “Understanding choice overload in recommender systems”. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 63–70.

[6] Nathan N Cheek and Andrew Ward. “When choice is a double-edged sword: Understanding maximizers’ paradoxical experiences with choice”. In: *Personality and Individual Differences* 143 (2019), pp. 55–61.

[7] Nathan N. Cheek and Barry Schwartz. “On the meaning and measurement of maximization”. In: *Judgment and Decision Making* 11.2 (2016), pp. 126–146. DOI: 10.1017/S1930297500007257.

[8] John P. Chin, Virginia A. Diehl, and Kent L. Norman. “Development of an instrument measuring user satisfaction of the human-computer interface”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’88. Washington, D.C., USA: Association for Computing Machinery, 1988, pp. 213–218. ISBN: 0201142376. DOI: 10.1145/57167.57203. URL: https://doi.org/10.1145/57167.57203.

[9] Ludovik Coba et al. “Decision making strategies differ in the presence of collaborative explanations: two conjoint studies”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. ACM, Mar. 2019, pp. 291–302. DOI: 10.1145/3301275.3302304.

[10] Ludovik Coba et al. “Exploring Users’ Perception of Collaborative Explanation Styles”. In: *2018 IEEE 20th Conference on Business Informatics (CBI)*. Vol. 01. July 2018, pp. 70–78. DOI: 10.1109/CBI.2018.00017.

[11] F Comstock. In: *prompt engineering. Encyclopedia Britannica*. Mar. 3, 2026.

[12] Dev K. Dalal et al. “Understanding the Construct of Maximizing Tendency: A Theoretical and Empirical Evaluation”. In: *Journal of Behavioral Decision Making* 28.5 (2015), pp. 437–450. DOI: https://doi.org/10.1002/bdm.1859. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.1859. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.1859.

[13] Fabrizio Dell’Acqua et al. “Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality”. In: *Organization Science* 37.2 (2026), pp. 403–423.

[14] Dalia L. Diab, Michael A. Gillespie, and Scott Highhouse. “Are maximizers really unhappy? The measurement of maximizing tendency”. In: *Judgment and Decision Making* 3.5 (2008), pp. 364–370. DOI: 10.1017/S1930297500000383.

[15] Liat Ein-Dor et al. “Conversational Prompt Engineering”. In: (Aug. 2024). DOI: 10.48550/ARXIV.2408.04560. arXiv: 2408.04560 [cs.CL].

[16] Jie Gao et al. “A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration”. In: *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. ACM, May 2024, pp. 1–11. DOI: 10.1145/3613905.3650786.

[17] E Gregersen. In: *ChatGPT. Encyclopedia Britannica*. Oct. 17, 2025.

[18] Michael A. Hedderich et al. *What’s the Difference? Supporting Users in Identifying the Effects of Prompt and Model Changes Through Token Patterns*. 2025. DOI: 10.48550/ARXIV.2504.15815.

[19] Joshua Holstein, Moritz Diener, and Philipp Spitzer. *From Consumption to Collaboration: Measuring Interaction Patterns to Augment Human Cognition in Open-Ended Tasks*. 2025. DOI: 10.48550/ARXIV.2504.02780.

[20] Sheena S Iyengar, Rachael E Wells, and Barry Schwartz. “Doing better but feeling worse: Looking for the “best” job undermines satisfaction”. In: *Psychological Science* 17.2 (2006), pp. 143–150.

[21] Dietmar Jannach and Markus Zanker. “Value and impact of recommender systems”. In: *Recommender systems handbook*. Springer, 2012, pp. 519–546.

[22] Michael Jugovac, Ingrid Nunes, and Dietmar Jannach. “Investigating the Decision-Making Behavior of Maximizers and Satisficers in the Presence of Recommendations”. In: *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. UMAP ’18. ACM, July 2018, pp. 279–283. DOI: 10.1145/3209219.3209252.

[23] Carolin Kaiser et al. “A New Era of Online Search? A Large-Scale Study of User Behavior and Personal Preferences during Practical Search Tasks with Generative AI versus Traditional Search Engines”. In: *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. CHI EA ’25. ACM, Apr. 2025, pp. 1–7. DOI: 10.1145/3706599.3720123.

[24] Kaeun Kim. “Maximizers’ Reactance to Algorithm-Recommended Options: The Moderating Role of Autotelic vs. Instrumental Choices”. In: *Behavioral Sciences* 13.11 (Nov. 2023), p. 938. ISSN: 2076-328X. DOI: 10.3390/bs13110938.

[25] Linda Lai. “Maximizing without difficulty: A modified maximizing scale and its correlates”. In: *Judgment and Decision making* 5.3 (2010), pp. 164–175.

[26] SeoYoung Lee and Junho Choi. “Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity”. In: *International Journal of Human-Computer Studies* 103 (2017), pp. 95–105. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2017.02.005. URL: https://www.sciencedirect.com/science/article/pii/S1071581917300198.

[27] Yidong Liang et al. “How Users Interact with Generative Information Retrieval Systems: A Study of User Behavior and Search Experience”. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’25. ACM, July 2025, pp. 634–644. DOI: 10.1145/3726302.3729998.

[28] Do Xuan Long et al. *What Makes a Good Natural Language Prompt?* 2025. DOI: 10.48550/ARXIV.2506.06950.

[29] Nicolas Maechler et al. *Next Best Experience: How AI Can Power Every Customer Interaction*. Accessed: 21 May 2026. Oct. 2025. URL: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/next-best-experience-how-ai-can-power-every-customer-interaction.

[30] Kelong Mao et al. “Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023, pp. 1211–1225. DOI: 10.18653/v1/2023.findings-emnlp.86.

[31] M Mcdonough. In: *large language model. Encyclopedia Britannica*. Oct. 14, 2025.

[32] McKinsey & Company. *What is personalization?* Accessed: 21 May 2026. May 2023. URL: https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-personalization.

[33] Fengran Mo et al. “ConvGQR: Generative Query Reformulation for Conversational Search”. In: (May 2023). DOI: 10.48550/ARXIV.2305.15645. arXiv: 2305.15645 [cs.IR].

[34] Carol Moser et al. “No Such Thing as Too Much Chocolate: Evidence Against Choice Overload in E-Commerce”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. ACM, May 2017, pp. 4358–4369. DOI: 10.1145/3025453.3025778.

[35] Gergana Y. Nenkov et al. “A short form of the Maximization Scale: Factor structure, reliability and validity studies”. In: *Judgment and Decision Making* 3.5 (2008), pp. 371–388. DOI: 10.1017/S1930297500000395.

[36] Jonas Oppenlaender, Rhema Linder, and Johanna Silvennoinen. “Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering”. In: *International Journal of Human–Computer Interaction* 41.16 (Nov. 2024), pp. 10207–10229. ISSN: 1532-7590. DOI: 10.1080/10447318.2024.2431761.

[37] Peter Pirolli. “Cognitive models of human-information interaction”. In: *Handbook of applied cognition* (2007), pp. 443–470.

[38] Peter Pirolli and Sanjay Kairam. “A knowledge-tracing model of learning from a social tagging system”. In: *User Modeling and User-Adapted Interaction* 23.2–3 (Nov. 2012), pp. 139–168. ISSN: 1573-1391. DOI: 10.1007/s11257-012-9132-1.

[39] Sruti Srinivasa Ragavan and Mohammad Amin Alipour. “Revisiting Human Information Foraging: Adaptations for LLM-based Chatbots”. In: (June 2024). DOI: 10.48550/ARXIV.2406.04452. arXiv: 2406.04452 [cs.HC].

[40] Steven Randazzo et al. *GenAI as a Power Persuader: How Professionals Get Persuasion Bombed When They Attempt to Validate LLMs*. Working Paper 26-021. Harvard Business School, 2025.

[41] Laurens Rook, Adem Sabic, and Markus Zanker. “Engagement in proactive recommendations: The role of recommendation accuracy, information privacy concerns and personality traits”. In: *Journal of Intelligent Information Systems* 54.1 (Oct. 2018), pp. 79–100. ISSN: 1573-7675. DOI: 10.1007/s10844-018-0529-0.

[42] Laurens Rook, Markus Zanker, and Dietmar Jannach. “Are We Losing Interest in Context-Aware Recommender Systems?” In: *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. UMAP ’24. ACM, June 2024, pp. 229–230. DOI: 10.1145/3631700.3665190.

[43] Barry Schwartz et al. “Maximizing versus satisficing: happiness is a matter of choice.” In: *Journal of personality and social psychology* 83.5 (2002), p. 1178.

[44] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. “Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking”. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI ’24. ACM, May 2024, pp. 1–17. DOI: 10.1145/3613904.3642459.

[45] Dietrich Silber, Arvid Hoffmann, and Alex Belli. “Embracing AI advisors for making (complex) financial decisions: an experimental investigation of the role of a maximizing decision-making style”. In: *International Journal of Bank Marketing* 43.6 (Mar. 2025), pp. 1325–1346. ISSN: 1758-5937. DOI: 10.1108/ijbm-10-2024-0647.

[46] Herbert A Simon. “Rational choice and the structure of the environment.” In: *Psychological review* 63.2 (1956), p. 129.

[47] Herbert A Simon. “Rational decision making in business organizations”. In: *The American economic review* 69.4 (1979), pp. 493–513.

[48] Herbert A Simon et al. “Theories of bounded rationality”. In: *Decision and organization* 1.1 (1972), pp. 161–176.

[49] Herbert A. Simon. “A Behavioral Model of Rational Choice”. In: *The Quarterly Journal of Economics* 69.1 (Feb. 1955), p. 99. ISSN: 0033-5533. DOI: 10.2307/1884852.

[50] Sofia Eleni Spatharioti et al. “Effects of LLM-based Search on Decision Making: Speed, Accuracy, and Overreliance”. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. ACM, Apr. 2025, pp. 1–15. DOI: 10.1145/3706598.3714082.

[51] Johanne R. Trippas et al. “What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild”. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 2024. ACM, July 2024, pp. 2703–2707. DOI: 10.1145/3626772.3657914.

[52] Brandon M. Turner et al. “The Maximization Inventory”. In: *Judgment and Decision Making* 7.1 (Jan. 2012), pp. 48–60. ISSN: 1930-2975. DOI: 10.1017/s1930297500001820.

[53] Lenka Vargová, L’ubica Zibrínová, and Gabriel Baník. “The way of making choices: Maximizing and satisficing and its relationship to well-being, personality, and self-rumination”. In: *Judgment and Decision making* 15.5 (2020), pp. 798–806.

[54] A Volle. In: *search engine. Encyclopedia Britannica*. Oct. 5, 2025.

[55] Ben Wang et al. “Task Supportive and Personalized Human-Large Language Model Interaction: A User Study”. In: *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*. CHIIR ’24. ACM, Mar. 2024, pp. 370–375. DOI: 10.1145/3627508.3638344.

[56] Fanghua Ye et al. “Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting”. In: (Oct. 2023). DOI: 10.48550/ARXIV.2310.09716. arXiv: 2310.09716 [cs.HC].

[57] Qinyuan Ye et al. *Prompt Engineering a Prompt Engineer*. 2023. DOI: 10.48550/ARXIV.2311.05661.

[58] Chanwoong Yoon et al. “Ask Optimal Questions: Aligning Large Language Models with Retriever’s Preference in Conversation”. In: (Feb. 2024). DOI: 10.48550/ARXIV.2402.11827. arXiv: 2402.11827 [cs.IR].

[59] Markus Zanker, Laurens Rook, and Dietmar Jannach. “Measuring the impact of online personalisation: Past, present and future”. In: *International Journal of Human-Computer Studies* 131 (Nov. 2019), pp. 160–168. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2019.06.006.

[60] Jizhi Zhang et al. “Envisioning Recommendations on an LLM-Based Agent Platform”. In: *Communications of the ACM* 68.5 (2025), pp. 48–57.

A

Questionnaire Item Coding

Table A.1: Mapping of user satisfaction items to shortened variable names.

Short name	Full item
satisfaction_1	I was satisfied with the experience of using the search mechanism to complete tasks.
satisfaction_2	Interacting with the search mechanism was a pleasant and satisfactory experience.
satisfaction_3	The dialogue with the search mechanism gave me useful information.
satisfaction_4	The overall assessment of conversing with the search mechanism was satisfactory.

Table A.2: Mapping of goal items to shortened variable names.

Short name	Full item
goal_1	I don't like having to settle for good enough.
goal_2	I am a maximizer.
goal_3	No matter what I do, I have the highest standards for myself.
goal_4	I will wait for the best option, no matter how long it takes.
goal_5	I never settle for second best.
goal_6	I never settle.
goal_7	No matter what it takes, I always try to choose the best thing.

Table A.3: Mapping of strategy items to shortened variable names.

Short name	Full item
strat_1	I can't come to a decision unless I have carefully considered all of my options.
strat_2	I take time to read the whole menu when dining.
strat_3	I will usually continue shopping for an item until it reaches all of my criteria.
strat_4	I usually continue to search for an item until it reaches my expectations.
strat_5	When shopping, I plan on spending a lot of time looking for something.
strat_6	When shopping, if I can't find exactly what I'm looking for, I will continue to search for it.
strat_7	I find myself going to many different stores before finding the thing I want.
strat_8	When shopping for something, I don't mind spending several hours looking for it.
strat_9	I take the time to consider all alternatives before making a decision.
strat_10	When I see something I want, I always try to find the best deal before purchasing it.
strat_11	If a store doesn't have exactly what I'm shopping for, then I will go somewhere else.
strat_12	I just won't make a decision until I am comfortable with the process.

B

Manipulation Check

Table B.1: Manipulation check for responses across questionnaires.

ID	Assigned Tool	Personality	Task 1	Task 2	Demographics
1	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
2	LLM	LLM	LLM	LLM	LLM
3	LLM	LLM	LLM	LLM	LLM
4	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
5	LLM	LLM	LLM	LLM	LLM
6	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
7	LLM	LLM	LLM	LLM	LLM
8	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
9	LLM	LLM	LLM	LLM	LLM
10	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
11	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
12	LLM	LLM	LLM	LLM	LLM
13	LLM	LLM	LLM	LLM	LLM
14	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
15	LLM	LLM	LLM	LLM	LLM
16	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
17	LLM	LLM	LLM	LLM	LLM
18	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
19	LLM	LLM	LLM	LLM	LLM
20	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
21	LLM	LLM	LLM	LLM	LLM
22	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
23	LLM	LLM	LLM	LLM	LLM
24	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
25	LLM	LLM	LLM	LLM	LLM
26	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine
27	LLM	LLM	LLM	LLM	LLM
28	Search Engine	Search Engine	Search Engine	Search Engine	Search Engine

C

Representative Search Chronologies

Table C.1: Representative task 1 chronologies for the LLM and search engine conditions.

LLM condition

Step	Action type	Action value
1	LLM query	A student from the TPM faculty at TU Delft has gotten the ” the steven hohhendijk best thesis award 2019-2020 from the batavian society for experimental philosophy”. i need you to find this thesis for me. please look at the tu delft thesis repository to see if you can find the correct thesis
2	LLM query	i only have this information. who won the steven hoogendijk best thesis award 2019-2020 from the batavian society for experimental philosophy? does the prize exist?
3	LLM query	The winner of the Steven Hoogendijk best thesis award in 2019 (the relevant cycle covering 2019–2020 references) is Dr.ir. Norbert Kalb. give me a link to this thesis
4	Website visit	repository.tudelft.nl
5	LLM query	”Bataafsch Genootschap der Proefondervindelijke Wijsbegeerte.” is this the batavian society for experimental philosophy
6	LLM query	it has to be master thesis, not a phd. can you find it?

Search engine condition

Step	Action type	Action value
1	Search query	TU Delft dteven hoogendijk best thesis award 2019-2020 batavian society for experimental philosophy
2	Website visit	qutech.nl
3	Search query	Norbert Kalb
4	Search query	Norbert Kalb tpm tu delft
5	Search query	TU Delft dteven hoogendijk best thesis award 2019-2020 batavian society for experimental philosophy
6	Website visit	research.tudelft.nl
7	Search query	TU Delft TPM steven hoogendijk best thesis award 2019-2020 batavian society for experimental philosophy
8	Website visit	repository.tudelft.nl
9	Search query	Norbert kalb thesis TU Delft TPM steven hoogendijk best thesis award 2019-2020 batavian society for experimental philosophy
10	Website visit	quantaneo.com
11	Search query	Norbert kalb thesis TU Delft
12	Search query	Diamond-based quantum networks with multi-qubit nodes
13	Website visit	link.aps.org
14	Website visit	arxiv.org
15	Search query	scholar
16	Website visit	scholar.google.com
17	Website visit	catalogue.leidenuniv.nl

Table C.2: Representative task 2 chronologies for the LLM and search engine conditions.

LLM condition

Step	Action type	Action value
1	LLM query	what is the condition of this apple
2	Website visit	hunker.com
3	LLM query	why do you think it is bitter rot
4	LLM query	please dubble check if you think this is the right condition of this appple
5	LLM query	what is going on with this apple, tell me what the conditions is of this apple please
6	LLM query	so both of the apples have the same conditoin ?

Search engine condition

Step	Action type	Action value
1	Search query	fruit going bad stages wikipedia
2	Search query	stages of ripening
3	Search query	stages of ripening apple
4	Search query	stages of over ripening apple
5	Search query	stages of over stale apple
6	Website visit	thishealthytable.com
7	Search query	stages of molding apple
8	Website visit	canr.msu.edu
9	Website visit	extension.psu.edu
10	Website visit	theorchardmaster.in
11	Website visit	plantscraze.com
12	Search query	stages of decay apple
13	Website visit	decompositiontime.com
14	Website visit	meatcheftools.com
15	Website visit	patentedbynature.wordpress.com