

Investigating How Humans Will Interface with an LLM-Integrated Internet

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Roy Jad

Summer 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Joshua Earle, Department of Engineering and Society

Introduction

Yesterday, I used a chatbot called ChatGPT to quickly and easily explain to my 5-year-old cousin why the sky is blue. Instead of sifting through page results — I simply typed “Explain why the sky is blue like I’m 5 years old. Come up with analogies to replace the concepts that could be hard to grasp.” ChatGPT, which is powered by Large Language Models (LLMs), conversed with me seemingly effortlessly as I continued to ask follow-up questions from the five-year-old. It initially generated two options:

Imagine you have a big box of crayons, and when you shine a light through the box, the blue crayon shines the brightest and makes it through while the others get lost inside the box. That's what happens in the sky. The sunlight comes through the air, and the blue light shines the brightest, so the sky looks blue to us.

The sky is blue because the sunlight likes to play with the air, and the blue light loves to dance around the most. So, when we look up, we see the blue light dancing in the sky, and that's why the sky is blue!

Let me know if that fits what you're looking for!

The caveat here is that I had to fact-check the answers I was given since the LLM within ChatGPT doesn't cite sources or scrape the internet once inquired. The LLM synthesizes responses that make sense based on the inquiry and the training it has had on vast segments of the internet. Luckily, in the case of my little cousin, the responses proved sufficient with verification — we both left the conversation satisfied. While convenient, however, the direct nature of internet search with LLMs can be a cause for concern.

LLMs are a type of artificial intelligence (AI) trained on large corpora of text (normally, large swathes of the internet), which specialize in predicting the “next word” with very high accuracy given an initial prompt — this enables them to have human-like conversations. This can save users time in contrast to using search engines to comb the internet and includes other benefits such as enabling users to tailor their responses to specific contexts.

To some, LLMs may seem magical in the way that they work, or appear to be intelligent or sentient; however, the way that they function is based on statistical patterns. It is important to note that they do not have an understanding of what they are saying. Since LLMs often appear to be “confident” in their answers but have no inherent ability to identify correctness, they can easily mislead/misinform users (Bender et al., 2021).

LLMs have made strides in recent years in terms of their effectiveness, and so we are seeing them being widely used in the form of ChatGPT and other chatbots, as well as being integrated into current widely used tools such as Bing Search and Google Workspace. With the advent of these LLM-integrated tools, it becomes particularly important to investigate them, including the ways in which they can have positive or negative impacts on society.

Objective

In this paper, I explore how Large Language Models (LLMs) can elevate the convenience of internet search for the average user, while examining the risks of their use.

Methods

In this paper I employ two strategies within the Value Sensitive Design (VSD) framework to explore how Large Language Models (LLMs) can change how we search the internet. The VSD framework emphasizes the importance of human values throughout the design process, guiding the investigation.

I first conducted a literature review. I investigated the history of LLMs/chatbots and how they've evolved by looking into seminal academic papers as well as modern articles to construct a wide overview of the technology. Through this overview, I contextualize the current state of LLMs and how they may fit into our current tools. In an effort to present a balanced viewpoint on this technology, I sought viewpoints from both sides — from those who see the technology as dangerous or unproductive, to those who believe it will pave the way for everything we do in the future. This literature review, conducted through the lens of VSD, evaluates the capabilities and limitations of LLMs, identifying how they can either enrich or hinder the online human experience.

Second, I conducted an interview. I spoke with an internet user who is in Gen-X and uses the internet every day for work and personal tasks — she can be described as proficient. I remotely observed her screen as she completed search tasks with and without LLMs, and then I

inquired about her experience. This hands-on perspective, combined with the VSD framework, illuminates the values and considerations that users may hold important when utilizing LLMs.

Literature Review

In this literature review, I dive into the history of LLMs and LLM-like tools as well as their effectiveness and various applications. I strive to include equal coverage toward the concerns described by AI scholars and workers in the field.

Hosseini-Asl, E., et. al. (2020) discuss simple language models (conversational AI) oriented for tasks in their paper which was written a few years prior to the emergence of LLMs, providing insight as to how natural language interactions were simulated without the intelligence of the LLMs we have at our hands today. Their work can give context to the way conversational interfaces have been evolving and will continue to evolve with the emergence of new technologies. Before the rise of LLMs, simple tools were created to enable basic conversation between humans and computers. These tools used basic rules and patterns to respond to user queries but lacked the ability to understand and interact with complex language. The research by Hosseini-Asl et al. shows how early conversational models were able to carry out simple tasks like answering common questions or assisting with basic needs. However, these systems struggled when faced with more complicated requests or nuanced conversations.

What stands out in Hosseini-Asl's work is the contrast between these early efforts and today's advanced LLMs. While once a significant achievement, those initial tools now seem almost primitive compared to the capabilities of modern language models. This transition from basic conversation tools to sophisticated LLMs tells a story of technological growth and

adaptation. The path from the early work of researchers like Hosseini-Asl to today's cutting-edge technology highlights the rapid pace of innovation in the field. It paints a picture of how our ability to communicate with machines has evolved and continues to advance, offering a glimpse into a future where human-computer interactions may become even more seamless and intuitive.

Some of the literature highlights the potential benefits of LLM-like conversational interfaces. For instance, Papenmeier et al. demonstrate that a chatbot-inspired search interface can lead to more engaging and personalized interactions. With keyword search, users have to "adapt to the system's language to maximise search success," which often results in users withholding "information about their initial information need." This is the case with standard search methods such as search engines. Users have to craft an input that they think the Google or Bing search will work best with, and then sometimes use specific search operators to further refine what they mean (*Refine Google Searches - Google Search Help*, n.d.). This may work but requires users to input specific syntax rather than use natural language — which could feel more natural/easy. On the contrary, they found that "the chatbot interface elicits longer queries with a significantly higher number of key facts;" (Papenmeier, et al., 2021, p. 1) this was especially true after the user was posed with a follow-up question from the chatbot. This could imply that users feel more confidence engaging with chatbots in a human way, which could infer a more comfortable mode of user interaction. Additionally, with the stated increased information/context, the chatbot/LLM may be more effective and "understanding" of the user's intent. The intuitive feel of interaction coupled with the increased effectiveness of chatbot interfaces is conducive to the feeling of convenience from users.

LLM-enabled internet “search” interfaces can cater the content and tone of the response to the specific user using information provided on the user’s background and current situation. On July 20, 2023, OpenAI — one of the leading companies in the AI space and inventors of ChatGPT, which is one of the leading LLM chatbots with over 100 million monthly users (*ChatGPT Revenue and Usage Statistics (2023)*, 2023), unveiled a new feature called custom instructions. Custom instructions enable users to provide ChatGPT with context and constraints as a static pretext for each new conversation — saving users time and conveniently catering each response to the specific circumstance of the user.

While LLMs can seem incredibly convenient, their use can also pose challenges — one notable concern is the occurrence of hallucinations. These are instances where the models generate text that seems correct but may not be grounded in actual reality. LeCun, an expert in the field, cautions, “Large language models have no idea of the underlying reality that language describes” (Smith, 2023). In essence, they can respond to prompts with answers that are statistically consistent but may lack true understanding. This limitation stems from the fact that most human knowledge goes beyond language; it's rooted in common sense and real-world experience. LeCun argues, “There is a limit to how smart they can be and how accurate they can be because they have no experience of the real world” (Smith, 2023). LLMs like ChatGPT are trained on billions of words, and while they can represent words and concepts through a mechanism known as embedding, this doesn't compensate for their lack of experiential learning. Even though abstract ideas can be learned through text, the absence of direct observation and nonlinguistic knowledge may continue to hinder their accuracy. The issue of hallucinations in

LLMs underscores the need for a more nuanced understanding of how these tools function. Connecting them more firmly to real-world knowledge and experience could be essential in improving both their effectiveness and trustworthiness.

In addition to hallucinations, LLMs present the risk of inherent bias. Since they are trained on internet content, LLMs are susceptible to the harmful content within these datasets. Even with careful filtering of specific keywords from the dataset, there are hegemonic views still inherent in the data of the internet which could be hard to detect. This effect is likely magnified because the subsections of the internet that were scraped encompass specific demographics of users. In one instance, the model GPT-2 was trained “by scraping outbound links from Reddit, and Pew Internet Research’s 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29” (Bender et al., 2021, p. 4). Users who enjoy the convenience of LLMs but do not consider the potential biases that come with them risk welcoming and spreading harmful misinformation, so it is essential that we design these tools with careful consideration at each stage of the process.

In summary, the literature reviewed provides a multifaceted overview of the topic, supporting my hypothesis that LLM-integrated interfaces can elevate the experience of internet search while underscoring some aspects of the problematic nature of LLMs.

Interview

I conducted a short interview to supplement the literature review by incorporating a human perspective. The interview was set up to observe how an average internet user would compare the use of internet search with and without LLMs in completing everyday tasks. I

interviewed a user in Gen-X — who will be referred to as Joy — who represents the average internet user in that she searches the internet occasionally for work and for personal research. She hasn't interacted with advanced chatbots before — only limited support bots. To prepare, I informed Joy of LLMs and how they work. I chose Zoom for the live interview due to its widespread use and screen-sharing tools. Joy consented to the use of video and screen recording.

The interview consisted of the oversight of two tasks: searching the internet for specific information in a work-related scenario and in a personal scenario. For each task, Joy had to perform it twice: once without using LLMs and once with using LLMs. The search engine of choice was Bing due to its effective LLM-integrated search and the ability to easily toggle the LLM integration. I then asked her to rate the ease of task completion, the challenges she faced, and the confidence she had in the results on a scale of 1 to 5. I also asked her to share her overall impression of using LLMs for internet tasks and how they compared to her expectations.

Task 1: Looking up Congestive Heart Failure Symptoms

Without LLMs:

Joy found the task relatively straightforward, giving it a 4 out of 5 for ease of completion. While she was confident about the accuracy of the results, she did note some challenges due to the sheer volume of information, occasionally losing track of what she was looking at.

With LLMs:

When using LLMs, Joy's experience improved significantly. She gave a perfect score for ease of task completion and enjoyed the clarity and targeted information that the LLM provided. She didn't face any challenges, found the LLM to be a great aid, and was confident in the accuracy of the results.

Task 2: Finding a Tiramisu Recipe

Without LLMs:

Joy scored the ease of task completion as 4 out of 5 but encountered difficulties due to multiple sources. She found it challenging to keep track of the initial recipe among many results.

With LLMs:

When using LLMs, Joy found the task easy and understandable. She appreciated the sentence completion feature and the smaller, more manageable selection of recipes. She had no challenges and expressed complete confidence in the accuracy of the results.

Overall Impression:

Joy came away with a positive impression of using LLMs for internet tasks and plans to continue using them. She views LLMs' answers as accurate and helpful and feels that their performance exceeded her initial expectations. Joy enjoyed the experience but mentioned that she needs more familiarity with its usage.

I observed Joy's struggle to locate information with standard internet search, which seemed due to overwhelming information and confusing formats. I also noticed a significant amount of idle time and redundant actions (like repeatedly opening files) with the standard internet search. However, I was pleasantly surprised by her immediate understanding of the chatbot interaction, her prompt receipt of responses, and the positive change in her voice's tone. It became apparent that she found the LLM significantly easier and more efficient to use for her searches compared to traditional methods.

Conclusion

In conclusion, the advent of Large Language Models (LLMs) represents a significant advancement in the way we search and interact with the internet. By exploring the development, capabilities, benefits, and challenges of LLMs through a comprehensive literature review and a practical interview, this paper has offered insights into how LLMs are reshaping our online experiences.

The convenience and user-friendly nature of LLMs, as evidenced by both the literature and the personal experience of the interviewee Joy, show the substantial value of these tools in everyday internet searches. They provide a more intuitive interaction, enabling users to communicate in natural language rather than specific search syntax. This not only makes the searching process more engaging but also allows for more accurate and tailored responses, enhancing the overall user experience.

However, this paper also emphasizes that the use of LLMs is not without its risks and challenges. The phenomenon of hallucinations, where the model might generate ungrounded

responses, and the inherent biases in the training data, highlight the complexity of relying solely on LLMs. These limitations are an important reminder that while LLMs can simulate human-like conversations, they lack real-world understanding and experience, and we must be wary of this during their use.

The convenience offered by LLMs can be outweighed by the harm of their potential misinformation or misguidance, given their inability to verify the accuracy of their own responses. This underscores the need for users to be mindful of fact-checking and verification when using LLMs.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. ACM Digital Library. <https://doi.org/10.1145/3442188.3445922>

Curry, D. (2023, May 5). ChatGPT Revenue and Usage Statistics (2023). Business of Apps. <https://www.businessofapps.com/data/chatgpt-statistics/>

OpenAI. (2023, July 20). Custom instructions for ChatGPT. <https://openai.com/blog/custom-instructions-for-chatgpt#OpenAI>

Hosseini-Asl, E., McCann, B., Wu, C. S., Yavuz, S., & Socher, R. (2020). A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33, 20179-20191.

Papenmeier, A., Kern, D., Hienert, D., Sliwa, A., Aker, A., & Fuhr, N. (2021). Starting Conversations with Search Engines - Interfaces that Elicit Natural Language Queries. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (CHIIR '21)*, 261-265. <https://doi.org/10.1145/3406522.3446035>

Refine Google searches - Google Search Help. (n.d.).

<https://support.google.com/websearch/answer/2466433?hl=en>

Smith, C. S. (2023, March 29). ChatGPT's Hallucinations Could Keep It from Succeeding. *IEEE Spectrum*. <https://spectrum.ieee.org/ai-hallucination>