Evaluating The Affordances of Large Language Models and the Potential Benefits to Usability

Liam Houston
Michigan Technological University
lghousto@mtu.edu

Abstract

This paper examines the potential impact of large language models (LLMs) on human-computer interactions (HCI), particularly in enhancing usability and accessibility. With rapid advancements in LLM technology, the integration of natural language processing (NLP) capabilities into HCI systems can lead to more intuitive interfaces and improved user experiences. The paper discusses various affordances of LLMs, including predictive text capabilities, context awareness, real-time translation, and error detection. Ultimately, this paper highlights the dual nature of LLMs as both a promising tool for improving HCI and a technology that requires careful consideration and further research to understand its implications for usability effectively.

Transparency statement

Microsoft Copilot was used as a tool to locate articles with specific content matter for use in this paper. Prompts were given to the Copilot AI requesting peer reviewed articles with specific subject matter relating to the topic of research. Only articles that were determined pertinent have been included. There is no concern of hallucination or misinformation in regards to the content Copilot provided. Only a small portion of articles were sourced using this method, not the entirety of the sources used. This is the only way in which an AI tool was used in the writing of this paper. Copilot was not consulted for ideas on the subject matter. All written content and ideas were generated solely by human authorship.

1. Introduction

The relatively quick advancements in large language models (LLMs) in recent years will undoubtedly have a very significant impact on the field of human-computer interactions (HCI). The field of HCI is constantly on the quest to enhance the usability and accessibility of future systems. The field of HCI stands to gain a lot from utilizing LLMs and their capabilities to process natural human language towards this goal. By leveraging LLMs, researchers and developers can create more intuitive interfaces with built in LLMs that improve user experiences and solve specific user needs. This paper explores the unique capabilities of LLMs and how they can be applied to HCI to improve usability, focusing on their abilities to improve user experiences through their capability to process language and context, identify and correct errors, and generate textual outputs. Overall, a more complete understanding of the place LLMs have within the field of HCI is the intended outcome.

There is currently very little distinction between subfields of generative AI, which would distinguish LLMs from generative AIs that are composed of text, audio, and image generation functions. Only LLMs will be discussed in this paper, not generative AI as a whole.

This paper is not an analysis or discussion of the usability of AI interfaces themselves. This paper is also not a consideration of AI tools as design assistants to HCI designers. This paper is a review of the capabilities of large language models of AI and what benefits they might bring as tools to increase the usability and accessibility of systems when incorporated within them.

2. LLM Affordances

The most notable affordance of LLM models is natural language processing (NLP). While not all generative AI tools use NLP, all LLMs are primarily composed of NLP systems. NLP is the defining characteristic of LLMs and is what separates them from other machine learning and generative AI. NLP is what allows LLMs to communicate effectively with users through natural human language [4]. NLP is made up of two primary language model types: contextual language models (CLM) and predictive language models (PLM) [1].

CLMs use previous words to predict future words in a sequence. CLM training is heavily supervised and involves encoding extensive pairs of input contexts that result in the model's predictive capabilities [2]. PLMs are trained without supervision on a large corpus of text. PLMs gain next-word predictive capabilities from analyzing preceding words in text and are able to generate textual outputs [3].

NLP and the different language models used to make LLMs are responsible for most of the beneficial functions of LLMs and are responsible for several useful applications of LLMs, which will be covered below.

2.1. Predictive Text Capabilities

The methods by which LLMs are trained allow them the ability to accurately predict and suggest text. This has been applied in a few areas to make user assistance tools like Google's Smart Compose LLM [5]. These capabilities have also been applied to coding interfaces like in Microsoft's Copilot [6]. LLMs like these are able to analyze what text a user has already

begun to compose and can use their encoded context and word sequence predictions to offer composition advice to users. LLMs can also suggest entire sections of text based on pre-established requests or user prompts.

Predictive text features have been shown to improve writer efficiency in college students using Smart Compose. The automated spelling and grammar provided through text suggestions also reduced the student errors [7]. It is likely that similar efficiency benefits are afforded to systems that use predictive text features. Although the state of research is still relatively new, making it hard to draw definitive conclusions.

2.2. Context Awareness

LLMs have become incredibly proficient at understanding context within a text. LLMs achieve this ability through two major components. LLMs separate sets of lingual data into linguistic units, represented by numerical labels within the LLM. These linguistic units have several levels, starting at the character-level and ending with phrase-level units [1]. These units—also referred to as tokens—are what allow LLMs to assign meanings to words and understand semantic information.

Improved search tools. The ability to encode semantic and contextual information through tokens allows LLMs to be used as more user-optimized search tools. LLMs that have been trained on a large set of data from across the internet are able to use the semantic and contextual information present within user input to find results more accurate to the user intent [1]. OpenAI's ChatGPT has shown itself to be beneficial at improving search efficiency. However, the quality of prompts can affect the usefulness of its results, and—like all LLMs—it is great at producing dictionary-level knowledge but not actual wisdom [8].

Abstractive text summary. When the system of tokenized information is combined with deep learning methods, a LLM also becomes fairly good at summarizing text to be better understood and quickly consumed by users. There are many algorithms made to accurately summarize text information, all of which have comparable results in the quality of results [9].

AI chatbots. A separate application of the same training and deep learning techniques allows AI to interact conversationally with users. Research into the effects of AI chatbots on usability has found that chatbots do have a positive impact on usability [10]. Chatbots are able to increase user engagement and guide users through using an interface, reducing user error and providing help when users get stuck.

Research into the use of AI chatbots on usability has recently been considered and proposed in the field of healthcare. Researchers from Saudi Arabia designed and intend to test the use of an AI chatbot as an assistance tool for aging adults to better access and use technologies related to their care [15]. The proposed research expects that declining eyesight and fine motor skills that affect patient access to technology can be accommodated through the use of a ChatGPT-3 based chatbot.

2.3. Real-time Translation

The proficiency with which an LLM can process and understand language makes it useful as a translation tool when it is trained on two or more languages. The generative capabilities of

LLMs allow translations made by an LLM to be accessible to users with practically zero wait time. The accuracy of LLM translation has been found to be incredibly reliable at reducing user error through its ability to provide real-time feedback and self correct errors [11].

While the translation capabilities of LLMs are focused on providing the highest quality output with as little latency as possible, there is a lack of research into whether the experience of users is actually affected by the use of LLMs over other translation programs and how automatic translations compare to human interpretations [12].

Whether LLM automatic translation is robust enough to replace other methods entirely is a major consideration facing HCI professionals. Although this translation capability is useful, it is unlikely that real-time translations made by LLMs will supplant translations generated by professionals in critical areas. Real time translation by NLP models is still heavily prone to error and cannot be fully trusted to return accurate outputs. Despite this, the efficiency they bring to the table makes them incredibly useful, even more so if current issues are sufficiently addressed [11].

2.4. Error Detection and Correction

The error detection and correction capabilities of LLMs are similar to their predictive text capabilities. LLM error detection uses the same CLM and PLM tools to analyze text, but instead of focusing on predicting the next words in the sequence LLMs check for and identify discrepancies between the input and the correct spelling or grammar within their training data [13]. Error detection can be applied as a tool to improve user efficiency and reduce error and cognitive load in the same way as predictive text services. LLM error detection capabilities have also been applied upon themselves to help improve the quality of user assistance chatbots. This is often achieved through a technique called reinforcement learning [14].

As mentioned earlier, there are already several commercial LLMs that utilize error detection and correction—as well as text predictions—to provide assistance to users [5, 13]. While not all LLMs are designed with writing checks or code assistance in mind, many popular AI services—ChatGPT, Copilot, Gemini, etc.—are capable of these functions if prompted.

3. Limitations and Concerns

3.1. Performance Limitations

Currently, there are a few notable performance limitations associated with LLMs that must be taken into account, as they are likely to affect their effectiveness as tools for enhancing usability. These limitations, such as hallucinations, biases in training data, and effects of limited training data, are bound to impact their effectiveness as tools to increase usability.

Hallucination. A persistent issue with LLMs is their propensity to produce convincing results that do not align with empirical realities. These outputs are referred to as hallucinations. LLMs can produce several types of hallucinations, ranging from factually incorrect fabrications to logical hallucinations that lack internal consistency and contextual hallucinations that lack consistent context to the original query [17].

The issue of hallucination can be made worse if the content in play is highly specialized. Research has shown that LLMs struggle to generate factual results, especially in cases where highly specialized knowledge is involved [8]. Since LLMs cannot truly understand the content they process, they are incapable of recognizing when their summaries are based on understandable language and not an accurate representation of the facts they were asked to interpret.

Group Bias. Bias towards certain groups or the promotion of certain biased viewpoints is a recurring issue for LLMs. LLMs are susceptible to pre-existing societal biases that are represented in some way within their original training data. Additional technical bias can be introduced to a system through the data input to the system [19]. The relationship between an LLM and user interactions can also unintentionally introduce bias. Users can impose their own bias upon the system through their inputs, which the system may learn to favor over time [20].

Training Limitations. LLMs are significantly limited by the quality of their training data. The data used to train an LLM dictated how that model behaves. The quality of training data is what can result in bias built into the model. Limitations in the scope of the data's knowledge can also lead to issues where LLMs cannot produce correct answers to requests relating to specific knowledge bases not within or well understood by the model. Since LLMs cannot actually understand what they know, a lack of information can often lead to mistakes in semantic encoding [21].

3.2. Other Challenges and Concerns

In addition to the purely functional limitations of LLMs, there are significant ethical and economic concerns that are important to the discussion of this technology. Considering these concerns is important if responsible implementation of LLMs on a larger scale is to be achieved.

Cost of LLMs. A major concern of LLMs is their high energy demands and the potential ramifications that may have on climate change. According to a study of the energy and emissions costs of several of the top task-specific and multi-purpose AI models, global data center electricity consumption has grown annually by 20-40% over the past few years. Training AI models takes the highest energy cost when, and due to the larger amount of training data, multi-purpose models tend to require more energy to train [22]. This study also found that comparatively, generative tasks cost more energy than discriminative ones, and task-specific models are more cost effective than generalized models overall. When it comes to the type of content being produced, image generation has much higher energy requirements than text generation.

Ethical concerns. The first major ethical concern when it comes to LLMs—and AI at large—is trust. Whether or not AI can be trusted is a very large question. AI is very good at generating believable outputs, something that causes a lot of trouble when believable outputs are factually misleading. Interestingly, a study into trust between humans and AI found that text-to-text interactions fostered higher trust in AI than text-to-image models. It also found that the same interaction effect that influenced trust also increased perceptions of usability [23].

The second major ethical concern with LLMs is transparency. Many individuals have valid concerns when it comes to AI transparency. AIs are not very good at showing users what's behind the curtain, leaving many users to wonder how AI has arrived at the outputs it gives them.

This lack of understanding can lead to distrust in the accuracy of AI outputs in some cases. There is also a lot of concern when it comes to individuals disclosing the use of AI to achieve their outcomes. Another area where the lack of transparency can lead to distrust [24].

4. Future prospects and the need for specialized models

The capabilities to improve system usability and accessibility that LLMs have make them a definite tool in the future of HCI. As previous sections have illustrated, there are many affordances of LLMs for usability applications and some promising research into practical applications of these affordances. As the tool of LLMs becomes more popular, the critical issues and limitations of the technology will have to be adequately addressed. The best way for HCI professionals to do this is through the development and testing of task-specific models as opposed to using commercial multi-purpose models.

Many of the aforementioned limitations of LLMs are mitigated by the creation of task-specific models. Most commercial generative AIs are generalized models intended for widespread use by a large and generalized user base. When applying these generalized models to specific use cases, they often underperform. Recent research would suggest that a lot of the performance issues with LLMs can be resolved through the development of task-specific models focused only on a small set of specific use cases [25]. As an example, other cases have found that task-specific models can reduce—or resolve entirely—the issue of hallucination purely through optimizing current model frameworks and closely monitoring model design and training [16, 18].

Currently, task-specific LLMs are also more cost effective to train and have the potential to work more efficiently than multi-purpose models. This and the performance improvements make small, task-specific models a better choice for HCI developers. As AI development becomes more accessible and cost effective it is likely that these benefits of task-specific models will continue to outweigh the convenience of using a commercially available generalized model.

Research into the use of AI in the field of HCI is still relatively new. Consequently, there is not a lot of research focused on the impact of LLMs—or any other AI—on usability. The state of research is still in its infancy, with the research that does exist on the topic being relatively recent. As the field of AI in HCI continues to evolve, it is essential for researchers and developers to prioritize the development of specialized models, fostering advancements that enhance user experience and accessibility across diverse applications.

5. Conclusion

The decision to implement LLMs as a usability tool remains a subjective choice, dependent on the specific needs and goals of the HCI designers and the systems they aim to enhance. It is apparent that there are several persistent concerns with LLMs as a technology at large. While LLMs offer considerable affordances for certain applications, their high cost and the need for specialized models to achieve optimal results make them a substantial investment. More research is needed to understand the impact LLMs have as a tool for usability and to identify methods of application that work best for HCI.

Given these considerations, it is not advisable to universally promote the use of LLMs within the HCI field as a method of improving usability. Although there are scenarios where LLMs

integrated into an interface can significantly improve usability, their effectiveness in this role should be better understood through additional research. The numerous limitations and other concerns for LLMs need to be better addressed in practical applications before the widespread utilization of LLMs as tools to improve user experience can be advised.

Sources

[1] Kumar, P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev 57, 260 (2024). https://doi.org/10.1007/s10462-024-10888-y

[2] Yu, F. H., Chen, K. Y., & Lu, K. H. (2022). Non-autoregressive asr modeling using pre-trained language models for chinese speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1474-1482.

[3] Wei C, Wang YC, Wang B, Kuo CC (2023) An overview on language models: recent developments and outlook. arXiv preprint arXiv:2303.05759

[4] Stryker, C., & Holdsworth, J. (2025, January 24). What is NLP (Natural Language Processing)?. IBM. https://www.ibm.com/think/topics/natural-language-processing

[5] Gmail smart compose: Real-time assisted writing. Google Research. (n.d.). https://research.google/pubs/gmail-smart-compose-real-time-assisted-writing/#:~:text=Abstract,currently%20being%20served%20in%20Gmail.

[6] WilliamDAssafMSFT. (n.d.). How to: Use copilot code completion for Fabric Data Warehouse - Microsoft Fabric. How To: Use Copilot Code Completion for Fabric Data Warehouse - Microsoft Fabric | Microsoft Learn. https://learn.microsoft.com/en-us/fabric/data-warehouse/copilot-code-completion

[7] Gnacek, M., Doran, E., Bommer, S., & Appiah-Kubi, P. (2020). The effectiveness of smart compose: An artificial intelligent system. Journal of Management & Engineering Integration, 13(1), 111-121.

[8] Cu, T. (2023). The Power of AI-Enhanced Search: Some Discussions on Its Benefits, Limitations and Bias. International Journal of Intelligent Information Systems, 12(3), 39-48. https://doi.org/10.11648/j.ijiis.20231203.11

[9] Saiyyad MM, Patil NN. Text Summarization Using Deep Learning Techniques: A Review. Engineering Proceedings. 2023; 59(1):194. https://doi.org/10.3390/engproc2023059194

[10] Dobbala, M. K., Lingolu, M. S. S. (2024). Conversational AI and Chatbots: Enhancing User Experience on Websites. American Journal of Computer Science and Technology, 7(3), 62-70. https://doi.org/10.11648/j.ajcst.20240703.11

[11] Mohammed, S. Y., & Aljanabi, M. (2024). Advancing Translation Quality Assessment: Integrating AI models for real-time feedback. EDRAAK, 2024, 1–7. https://doi.org/10.70470/edraak/2024/001

[12] Papi, S., Polak, P., Bojar, O.ˇ., & Mach'avcek, D. (2024). How "Real" is Your Real-Time Simultaneous Speech-to-Text Translation System? ArXiv, abs/2412.18495.

[13] Microsoft 365. (2024, March 29). How to use AI to help you improve your grammar. https://www.microsoft.com/en-us/microsoft-365-life-hacks/writing/how-to-use-ai-to-help-you-improve-your-grammar#:~:text=AI%20tools%20leverage%20sophisticated%20language,%2C%20punctuation%2C%20and%20contextual%20nuances.

[14] Izadi S, Forouzanfar M. Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots. AI. 2024; 5(2):803-841. https://doi.org/10.3390/ai5020041
[15] Khamaj, A. (2025a). Ai-enhanced chatbot for improving healthcare usability and accessibility for older adults. Alexandria Engineering Journal, 116, 202–213. https://doi.org/10.1016/j.aej.2024.12.090
[16] Ganguli, D., Askell, A., Schiefer, N., Liao, T., Lukovsiut.e, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., Drain, D., Li, D., Tran-Johnson, E., Perez, E., Kernion, J., Kerr, J., Mueller, J., Landau, J.D., Ndousse, K., Nguyen, K., Lovitt, L., Sellitto, M., Elhage, N., Mercado, N., Dassarma, N., Lasenby, R., Larson, R., Ringer, S., Kundu, S., Kadavath, S., Johnston, S., Kravec, S., Showk, S.E., Lanham, T., Telleen-Lawton, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T.B., Olah, C., Clark, J., Bowman, S., & Kaplan, J. (2023). The Capacity for Moral Self-Correction in Large Language Models. ArXiv, abs/2302.07459.
[17] Kaur Sidhu, B. (2025). Hallucinations in artificial intelligence: Origins, detection, and mitigation. International Journal of Science and Research (IJSR), 14(1), 8–15. https://doi.org/10.21275/sr241229170309
[18] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z., Dassarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T.B., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., & Kaplan, J. (2022). Language Models (Mostly) Know What They Know. ArXiv, abs/2207.05221.
[19] Stoyanovich, J., Howe, B., & Jagadish, H. V. (2020). Responsible data management. Proceedings of the VLDB Endowment, 13(12), 3474–3488. https://doi.org/10.14778/3415478.3415570
[20] Mehrabi, N., Morstatter, F., Saxena, N.A., Lerman, K., & Galstyan, A.G. (2019). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys (CSUR), 54, 1 - 35.
[21] Christine Cuskley, Rebecca Woods, Molly Flaherty; The Limitations of Large Language Models for Understanding Human Language and Cognition. Open Mind 2024; 8 1058–1083. doi: https://doi.org/10.1162/opmi_a_00160
[22] Luccioni, S., Jernite, Y., & Strubell, E. (2023). Power Hungry Processing: Watts Driving the Cost of AI Deployment? Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.
[23] Ding, Z., & Chiu, M. (2024). Can I collaborate with you? an investigation of trust in generative ai. Proceedings of the Association for Information Science and Technology, 61(1), 889–891. https://doi.org/10.1002/pra2.1130
[24] Rehill, P., & Biddle, N. (2024). Transparency challenges in policy evaluation with causal machine learning: Improving usability and Accountability. Data & Policy, 6. https://doi.org/10.1017/dap.2024.35
[25] Arash Hajikhani, Carolyn Cole; A critical review of large language models: Sensitivity, bias, and the path toward specialized AI. Quantitative Science Studies 2024; 5 (3): 736–756. doi: https://doi.org/10.1162/qss_a_00310