An Assessment of the Usability of ChatGPT

Johnathon Hall *jhall28@oswego.edu*
Department of Computer Science
State University of New York
Oswego, New York, 13126, USA

Damian Schofield *damian.schofield@oswego.edu*
Department of Computer Science
State University of New York
Oswego, New York, 13126, USA

Abstract

The prevalence of Artificial Intelligence (AI), in particular Large Language Models (LLMs) in multiple areas of society is rapidly growing. This surge in popularity has attracted users of all age groups, leading to a substantial increase in the number of individuals interacting with AI tools. This research aims to examine the way both current and new users engage with ChatGPT (a generative AI chatbot developed by OpenAI and launched in 2022) specifically for academic purposes, and evaluate the effectiveness of this engagement. The project seeks to scrutinize the accuracy of the results produced by ChatGPT, as well as the functionality of the interface and prompt generation within ChatGPT. Additionally, concerns regarding the ethical implications of employing an AI agent for academic research and writing along with accessibility and availability are examined. The study involves college students specializing in the field of biological science as well as their use of ChatGPT to develop research reports. This study finds that despite advances in ChatGPT, users struggle with creating effective inputs due to user interface challenges. It calls for improved LLM interfaces and user education while emphasizing the need for equitable access and ethical considerations by developers.

Keywords: Artificial Intelligence, Large Language Models, ChatGPT, Human-Computer Interaction, Usability, Ethical Implications, Biological Science.

1. INTRODUCTION

User Experience (UX) is a crucial area of Human-Computer Interaction (HCI) for technological and product advancement. UX frequently stands out as a requirement for successful software development endeavors. This is particularly true when it comes to novel technologies and previously unexplored user demographics (Martinelli et al., 2024; Mortazavi et al., 2024). UX now plays a significant role in various industries worldwide. However, compared to other fields, the field of AI is relatively new and lacks the same level of comprehension and expertise that we see in other fields (Kantorovitch et al., 2017).

Current UX research is discussed in many different ways concerning different types and forms of AI, past research has identified multiple key challenges in designing for usability with AI (Yang et al., 2020). Much of the past research has focused on the roles in which UX can be improved as well as how UX can improve AI, working both ways towards discovering potential improvements in methodologies that UX professionals apply in their line of work.

Furthermore, past research has also emphasized the insufficient attention given by UX professionals to prototyping with AI, though some of this can be attributed to the ongoing and evolving nature of AI development (Stige et al., 2024; Shah et al, 2022). The rapid growth of AI and the evolving nature of the workplace are significant factors in shaping usability requirements.

While tying into a later area of this literature review as well, the complexities of rapid test prototyping and how they impact user experience including that of societal measures are often claimed to be unable to be completed meaningfully (Xu et al., 2023). Other hindrances are also noted throughout the past research regarding the challenges for many UX leaders to understand the full capabilities of AI and a lack of technical experts in the field to collaborate with (Yang et al., 2020).

With regard to UX research examining metadata for upcoming projects;, evaluation outputs, and the creation of AI processes with which to use are all areas of thematic focus. Notably, this raises the issue of the definition of AI in relation to HCI/UX, which remains varied in the current body of literature (Liu et. Al., 2023; Xu et al., 2024; Lu et al., 2024). Some individuals go so far as to posit that ChatGPT is considered intelligent while others specifically have noted that it simply searches and finds information related to prompts and queries (Hanna, 2023).

The definition to be used in this paper is one from Yang et al. (2020) :

> *"AI in relation to computational mechanisms that interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation."*

The primary findings of Yang et al. (2020) challenge existing research that attributes the difficulties in designing and working with AI to its unpredictability. Their belief is that the root challenge of mapping rule-based interactions, touch on the nature of complexity in outputs by framing AI systems as living, sociotechnical systems which then offers a unique insight into how one could analyze, interpret, and design for AI systems on a whole.

Other studies mirror a sentiment that AI usability should be at the forefront of design and that a lag in this form of UX knowledge is occurring (Yang, Wei, & Pu, 2020).This research builds upon current/working HCI models for UX but focuses more on the design itself, they argue that UX designers lack knowledge of AI techniques. While this work explores designing for usability, this leads to an important question within the UX field. Much of the output generated from AI systems requires some form of input from the user's side, the how can designers and users assess the accuracy of what is returned as the output.

There is currently a focus on the need for bracketing AI within the collaboration and oversight areas, to make sure that AI is being utilized in meaningful and appropriate ways across mainly enterprise/business users (Hanses & Wang, 2022; Kuang et al., 2024). A particular challenge worth highlighting here is that product teams tasked with the design of AI elements have noted usability challenge for users in multiple phases of the user journey (specifically onboard, use, and maintain phases). These phases are important as they directly involve the input and output of AI systems. This is particularly evident in the case of chatbots like ChatGPT.

Being able to immediately see the results of an input results in easier understanding and could also lead to developmental improvements for HCI research. It is important that any improvements in usability are implemented so as to allow for new users who are intimidated by the processes of using AI (Baek and Kim, 2023).

Some texts focus on how educators can utilize the system for developing resources (such as courses, quizzes, and materials) using an AI system. Skrabut (2023) provides a comprehensive overview of 80 different educational applications of ChatGPT, and although there isn't so much of a question on the ability of an educator to use AI technology, there is perhaps a question regarding the ethics of doing so. However, the author does acknowledge the importance of educating students about attributing credit to AI co-writers, which is currently a relevant topic of controversy in higher education and research writing (Lee et al., 2024).

Skrabut (2023) also provides a schematic for instructions on how to use ChatGPT. Simplified instructions for the use of AI systems, in educational settings, are perhaps something that should be readily available within the AI application themselves. These instructions should also be

offered as instructional components for educators and students alike when implementing these systems in the changing educational landscapes ahead. More importantly, this is an area where the UX can certainly be improved. There is little available research, and few case studies reported, in this area.

2. DETERMINING ACCURACY AND VALIDITY

One of the major concerns over ChatGPT has been the validity or accuracy of the claims and statements that the AI chatbot makes when answering any prompt. These statements themselves are difficult to check, and this checking often requires extra time and effort from the user (Lechien et al., 2024; Yilmaz et al., 2024).

The power of ChatGPT 3 has been proven by it's ability to pass the United States Medical Licensing Examination (USMLE). Kung et al. (2023) reported that the system could pass the examination with a 60% accuracy rate for the answers given and didn't require any form of specialized input. The questions asked were open-ended and multiple-choice and the way that questions are asked is such that there is only one right or wrong answer due to test constraints. Although, there is often difficulty in ascertaining the correct answer due to conceptually dense text vignettes that contain multimodal clinical data and are used to generate ambiguous scenarios. In this USMLE experiment, ChatGPT was forced to provide justifications of why the AI system chose its responses, in relation to the relevant level of medical knowledge. Many of the justification responses provided by ChatGPT gave back what were classed as non obvious insights, or unclear reasoning (Kung et al., 2023).

With the advent of GPT 4, there have been significant improvements in accuracy and nuanced scenarios (OpenAI, 2023). Currently, GPT 4 is performing in the top 10% of test takers who take the legal bar exam, with GPT 3.5 the scores were in the bottom 10% (Katz etal., 2024).

Many other standard exams have been tested using different versions of ChatGPT. While AI systems scored very highly on many exams (such as the GRE and AP science subjects), they performed very badly on the AP English Language, Literature and Composition exams (Rudolph et al., 2023; Afkarin & Asmara, 2024).

This deficiency in AI performance on these specific tests highlights the need for checking for validity and accuracy within results associated with these areas of knowledge. These exam results also invite a discussion on whether the validity of responses that AI systems, such as ChatGPT, will eventually reach a point of 100% accuracy, and be 100% verifiable.

Recently, there has been an instance of a judicial case being found to have used ChatGPT and given incorrect case citations (Alkaissi and McFarlane 2023). In this case, ChatGPT provided five references were given with regard to a question on "homocysteine-vitamin K-osteocalcin. None of the references provided were real papers. It is evident here that questions must be well-worded in order to elicit the desired responses that are correct, but also that additional fact-checking of information that is provided should become a priority for users who use these AI systems. This demonstrates the need for better usability of generative AI systems as a whole, with regard to accuracy and validity.

3. ETHICALITY AND SAFETY

While accuracy is a common area of generative AI research, another important area f research relates to the ethics of these generative systems. Ethical AI research is generally broken down into a number of common topic areas including: Bias, Robustness, Reliability, and Toxicity(Zhuo, Huang, Chen, & Xing, 2023; Vetter et al., 2024; Pant et al., 2024).

Research on individual ethics topics have identified a number of problems (Rahimi & Bezmin Abadi, 2023).Problems identified include stereotyping, discrimination, and exclusion of languages or peoples. This often occurs due to the mechanisms used to train LLMs. Although these are

among the more common ethical concerns, there are also issues with data leakage and reliability. Previous research, undertaken using GPT-3, showed a large amount of susceptibility to perturbations and a struggle to remove bias in the output of LLMs. The authors call for a larger understanding and definition of ethics within AI systems.(Zhuo, Huang, Chen, & Xing, 2023).

The issue of safety is always a prominent subject when discovering and discussing new technologies, AI and LLMs such as ChatGPT offer no exception to this rule. There are studies that have looked into the risks of particularly harmful information such as how to make a dangerous chemical at home using home ingredients and in layman's terms (OpenAI, 2023). GPT-4 is now able to identify that some of its output could be dangerous and redacts that information from the responses, noting to the user that it can't provide information on harmful substances.

A range of research has already been conducted on the safety of generative AI systems. However, there is still an issue that LLMs can potentially find enough information from a variety of different sources and modify their responses in the form of "propaganda" or incorrect answers with intentional meanings. Goldstein et al. (2023) investigated the stages of intervention required to stop propagandists. Four stages were noted for AI as particular areas of influence operations that would allow for bad actors to engage in and embark on large-scale propaganda:

- Model Design & Construction
- Model Access
- Content Dissemination
- Belief Formation

The last stage, belief formation, offers a mitigation factor of restrictions on usage as well as media literacy campaigns. These stages are very similar to those used in modern-day education for digital and web usage to teach students to think critically and examine the authenticity of researched items. The researchers note that there are likely to be improvements in usability, reliability, and efficiency of AI as time goes on and that these improvements will ultimately change the way individuals engage with these models and the way in which bad actors function. However, they also note that there is currently no 'silver bullet' for mitigating influence operations (Goldstein et al., 2023).

Other studies seemingly ignore the danger of forced misinformation and focus on the nature of ChatGPT failing the Turing test and its deference in giving an opinion on any particular topic (Noever & Ciolino, 2022).

A number of studies have examined LLMs and academic integrity, which is an increasing concern in the field of education. There has been a rapid increase in the number of articles on this issue in the literature (Currie, 2023). When considering the validity issues mentioned above, the question of the ethics of students using LLMs for academic fraud versus using them for the benefit of finding information is difficult to navigate and needs to be investigated further.

4. EXPERIMENTAL METHODOLOGY

A key component that is identified in the research discussed in the previous sections is the prompts that users entering into LLMs in order to elicit responses. It is evident that when asked a direct question, with a simple enough set of available responses that these systems will regularly perform well. However, there are many issues regarding the nature of more complex requests and a requirement to improve the usability for the users of generative AI systems in general.

With the improvement of this area of research, there will perhaps be a framework for guiding new users to safely, efficiently, and accurately use these tools. In addition, this will provide a step forward for UX specialists to begin setting up scaffolds for future and current improvements to generative AI systems.

4.1 Experimental Procedure

The research design is experimental, performed via screen share on the participant's own computer with the user creating an ChatGPT account (or using an account they already possessed). Once created, users were given a rubric for a Biology paper focused on biodiversity that contained a series of requirements and sections. The research assignment provided was in the field of Biological Sciences with a focus on research on the organism Pantherophis guttatus or the corn snake.

The participants then developed a prompt and entered this into ChatGPT and copied and pasted the output to answer the assignment. The participants then submitted the assignment to the research team. All files were converted to include no identifying markers and the original emails were deleted.

Many studies exist currently on a variety of different versions of ChatGPT. The gap between ChatGPT 3 and ChatGPT4 has been extensively discussed (Plevris et al., 2023; Sahib et al., 2023). The experimentation described within the study accompanying this literature review is using ChatGPT3 due to ChatGPT4 currently needing a subscription fee.

This study follows a deductive approach, starting with an existing theoretical framework and background knowledge about the different versions of ChatGPT (Kalla et al, 2023). The hypothesis concerns the performance or differences in output quality of ChatGPT3 (Urban et al, 2024; Lee et al (2024).

4.2 Experimental Participants

Participants were recruited from the summer biology lecture courses at the University of South Carolina. A total of 21 biology students participated in the experiment. There was no selection process or criteria that had to be met for participation other than being a student taking these courses. Prior to signing up for a time slot for screen share, participants were given an informed consent form to review. Of the 21 participants, 16 students participated in the full experiment with one response that was unable to be used due to file corruption.

4.3 Experimental Metrics

The participant papers were randomized with no prompt attached and sent to a professor of Biology at the University of South Carolina. This professor graded the papers based solely on the information required by a rubric (Table 1), and then submitted the papers through the plagiarism detection software at the university to determine a plagiarism rating.

The paper grade (points scored), prompt time, number of prompts used, and complexity of prompts were used to determine the overall effectiveness of the participants' use of ChatGPT to generate a usable research paper that satisfied the requirements of the rubric. All of the metrics selected were based on previous research in this field, which indicated areas that were of interest for future forward progress and overall experiences related to usage of LLMs. The papers were then checked using three different AI detectors to see if any trends were apparent.

5. EXPERIMENTAL RESULTS

Throughout the experiment, a number of measurements were taken for each experimental participant including :

- the grade (point value) against the provided rubric
- total prompt time
- total number of prompts used
- total number of characters used
- complexity of the prompts based on sentences
- percent of plagiarism detected
- percent of AI detected

Grading Rubric for Biodiversity Report

Criteria – Your Organism for this assignment is on Pantherophis guttatus	Points Possible	Points Earned
General Paper Information • Informative Title is included [2 point] _____	2	
Introduction • Interesting introduction about the organism [2 points] ____ • Identification of the organism [1 point] ____ • Correct use of scientific name (Capital Genus, lowercase specific epithet as well as entire name either italicized or underlined) [2 points] ____ • Brief synopsis of the primary literature article [3 points] ____	8	
Background • General information about the kingdom and/or subtaxa (number and type of cells, producer or consumer, reproduction, methods of defense, etc.) [5 points] ____ • General information about the species (habitat, food, interactions with other organisms, etc.) [5 points] ____ • Specific information about the species needed to understand the primary literature [3 points] ____ • Explanation of topics or concepts needed to understand the primary literature (define any terms that are necessary to understand the primary literature) [3 points]	16	
Data Analysis • Brief summary of the experiments carried out in the primary literature (explain what the scientists did in your own words) [5 points] ____ • Explanation of the results (explain the scientists' findings in your own words) [5 points] ____ • Inclusion & Explanation of 1 graph, table, or figure [5 points] ____ • Explanation about how the author(s) link the evidence to their results/conclusions [5 points] ____	20	
Conclusion • Statement about the organism in light of new evidence [2 points] ____ • Importance of the discovery [2 points] ____	4	
References • Minimum of two references (one primary and one secondary) [2 points] ____ • Proper format of references in bibliography [4 points] ____ • Proper use and format of in-text citations (since this is a literature review, **almost every sentence** should have an in-text citation…this does not mean that you should use direct quotes but **paraphrase** in your own words including in-text citations, see **PENALTIES** for point reductions)	6	
Writing Quality • Logical organization; clear train of thought and organization [2 points] ____ • Correct spelling and grammar (more than 5 mistakes will incur at least a point deduction) [1 point] ____ • Proper length and style requirements (3 page minimum, not including the Works Cited page) [1 points] ____	4	
Total	**60**	

The rubric detailed the sections required in the paper. Plagiarism was also not considered within the rubric as a detrimental point value due to the nature of this experiment, though it was measured according to multiple plagiarism and AI detectors. When checking against the AI detectors, the highest percentage was chosen in each case. The experimental results were then collated (Table 2).

TABLE 1: Grading Rubric for the biodiversity paper.

Parti c-ipant	Grade (Point Value)	Prompt Time	# of Prompts Used	# Of Character in Prompt	Complex Prompts ?	% of AI Detected	% of Plagiarism Detected
1	54	13:27	3	1147	Yes	77%	0%
2	53	12:22	3	1306	Yes	96%	0%
3	43	10:13	2	69	Yes	62%	0%
4	40	9:06	1	1620	Yes	71%	1%
5	38	1:35	1	100	No	70%	0%
6	31	2:56	2	198	Yes	75%	0%
7	31	4:23	2	3964	Yes	80%	0%
8	31	2:10	1	112	No	89%	0%
9	30	1:03	1	233	Yes	71%	0%
10	29	10:38	2	81	Yes	87%	0%
11	29	1:47	1	170	No	82%	0%
12	28	1:03	1	277	No	70%	4%
13	25	5:23	1	212	Yes	80%	0%
14	23	1:11	1	35	No	75%	0%
15	23	2:39	1	70	No	80%	0%

TABLE 2: Experimental metrics.

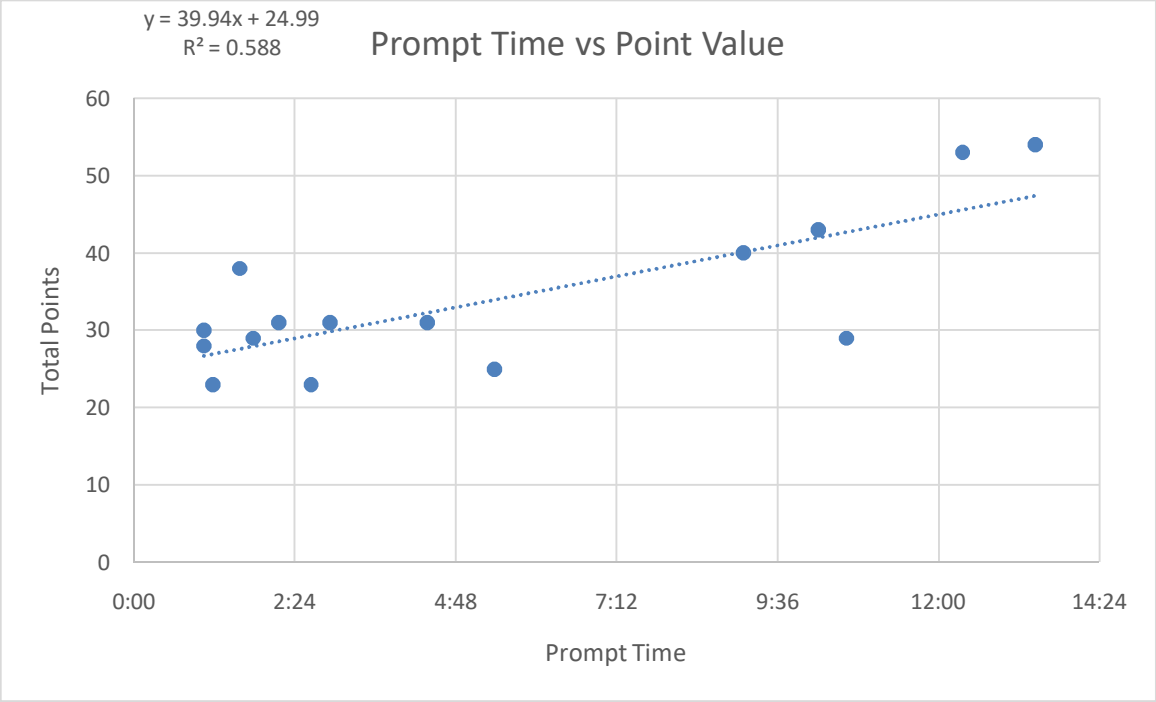


FIGURE 1: Prompt time vs grade (point value).

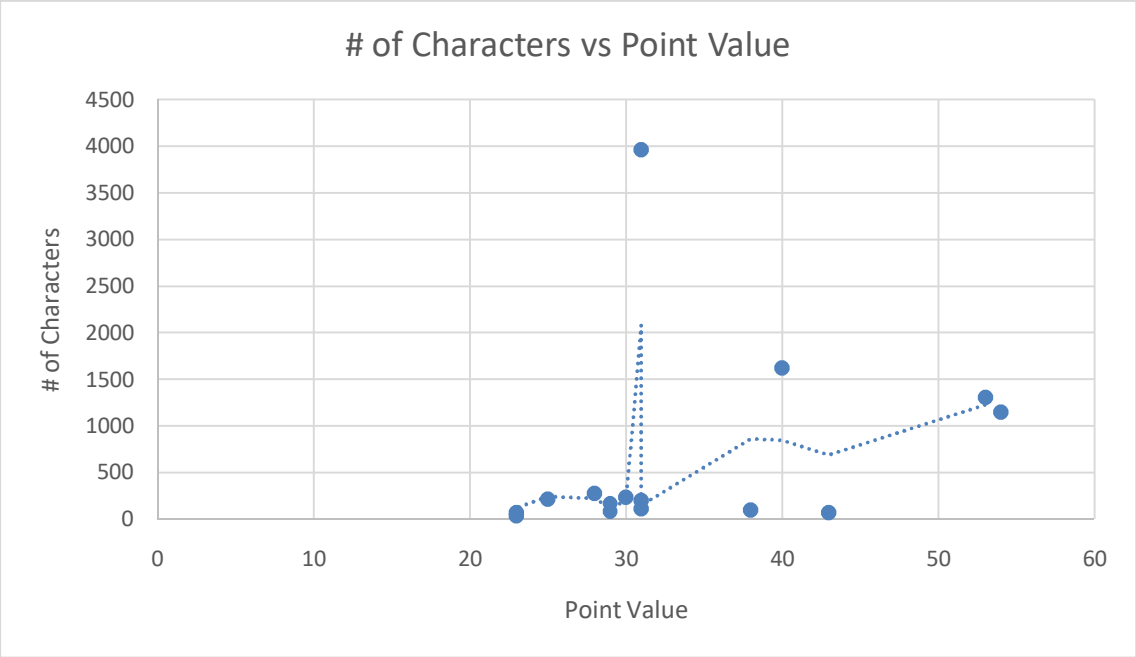
6. DATA ANLYSIS

A Pearson correlation was undertaken using prompt time and point values (Figure 1).The research involved a sample size (N) of 15 observations as in Table 2. The obtained t-statistic of 4.631 and the corresponding degrees of freedom of 13 indicate that the correlation coefficient is statistically significant between these two variables. The p-value of 0.0004 suggests that the obtained correlation coefficient is unlikely to have occurred by chance alone and is statistically significant with a Pearson correlation coefficient of 0.7671.

Overall, the research findings indicate a significant and positive relationship between the variables of prompt time and grade, as supported by the calculated Pearson correlation coefficient, t-statistic, degrees of freedom, and p-value,. This demonstrates that time spent developing the appropriate prompt positively correlated with higher point score potential.

An additional correlation test was done on the number of characters used to develop the prompts and the grade (Figure 2). The Pearson correlation coefficient of 0.2771 suggests a weak positive correlation between the variables being studied. The research involved a sample size (N) of 15 observations. The obtained t-statistic of 1.1172 and the corresponding degrees of freedom (df) of 13 indicate that the correlation coefficient is not statistically significant.

The p-value of 0.284099119 suggests that the observed correlation could reasonably occur by chance alone. With a p-value greater than the predetermined significance level (e.g., 0.05), there is not enough evidence to support the existence of a significant correlation between the variables of point value and number of characters used within the prompts. The lack of statistical significance and the relatively high p-value indicate that the observed correlation is not considered statistically significant and may be attributed to random variation in the data.



FIGUREZ: Number of characters n the prompt vs grade (point value).

7. DISCUSSION

Using the data from the Pearson correlations, as well as general scoring of the overall assessments, a few patterns emerge in three different areas.

- Time spent on a particular prompt often improves the output

- Complexity/size of the prompt does not always improve output
- The average student would fail to pass this assignment using ChatGPT

When reviewing the rubric and attempting to create/writepaper using an LLMs, the time spent on developing the prompt often improves the grade. The longer the participant spent on creating the prompt(s), the more likely they were to score higher. This provides an important potential recommendation for improving the user experience of users with LLMs. This is that improving the interface to help users choose and create prompts could be very useful, and greatly improve the quality of the output.

The second observation in the list above shows that the complexity of the prompt does not affect the overall result positively or negatively. When using the LLM in this experiment, participants with larger character amounts did not perform significantly better than their peers who used character counts of smaller sizes. Again, improving the overall usability of an LLM could potentially reduce both time and effort of the user while also improving the experience and quality of the output. This reinforces the conclusion that understanding of AI and LLM systems is not innately understood by users, but more of a learned process.

The final outcome illustrated by the data analysis is related to the general mean, median, and mode of the overall grade (point value) scored on the assignment. With a mean of 33.13, a median of 31, and a mode of 31, the majority of the students who completed this task failed the research paper assignment. The student participants in this experiment clearly understood how to write a research paper and how to follow the rubric, but still struggled to generate appropriate prompts for ChatGPT.

There were a number of examples of extremely detailed prompts, showing that the participants were trying to wrangle the AI system to give them the output they needed. Again, this problem illustrates the issue of how usable ChatGPT was for participants as they attempted to use it to write their papers. The interface was clearly, not very effective for the participants, and the quality of work generated by the LLM was often subpar. It is interesting to note that most of the work produced was actually grammatically sound but failed to meet basic requirements as set out by the rubric.

Finally it is worth mentioning the AI detection soft used in this study. All of the papers generated by the participants using ChatGPT were checked using three different AI detection tools:

- Copyleaks (www.copyleaks.com)
- Smallseotools (www.smallseotools.com)
- Plagiarism Detector (www.plagiarismdetector.net)

It is interesting to note that the Copyleaks system repeatedly showed much higher percentages of AI generated content in the participant's papers than the other two systems. This finding illustrates the different efficacy of the AI detection tools available and offers questions of continued resiliency to detect work that is entirely created by AI.

Considering the variability of the AI detection software, and the failure of some of the tools to detect generated AI content, it is obvious that it currently remains difficult to detect if a student has used an LLM to generate content. The current consensus is that educators of all levels should use extreme caution when stating that a student has definitely used an LLM to generate submitted work (Dhaini et al., 2023; Chaka, 2023).

8. CONCLUSIONS

While ChatGPT and other LLMs are continuously improving, in a similar manner to the advent of a search engine, asking the right questions can elicit the appropriate responses. However, as this experiment demonstrates that there is a large gap between the technical mechanism of

generating outputs versus the human side of creating appropriate inputs to the LLM. For the average user, generating content regarding any more complex topic appears to still be a difficult task when using an LLM.

Most users do not naturally understand how to converse with a generative AI system such as ChatGPT. They often fail to understand or redefine parameters as necessary when receiving responses that they believe are incorrect or insufficient resulting in sub-quality output.

There have been significant improvements in accuracy and nuanced scenarios with LLMs, and they have proved themselves by passing a number of complex tests and exams (OpenAI, 2023; Kung et al., 2023).

In the experiment described in this paper, the failure of ChatGPT to pass this assessment was not due to the failure of the technology as a whole, but due to a failure of the user interface between the experimental participants and the software itself. Most previous research ignores this important component of the efficacy of LLMs.

It is vital that UX specialists work to improve these LLM interfaces. This work should be completed early on in this phase of adoption as AI and LLMs become pervasive and ubiquitous in society. If UX designers and researchers do not take this opportunity to have input in these initial designs many users will be left behind as the system complexity increases. It is also the responsibility of UX specialists to ensure ethical transactions and accessibility for all members of society to be able to use these systems appropriately.

It is vital that time is devoted to prompt generation features, the onboarding, and the education of new users. Technology education is an area that schools of all ages begin to engage with at this time so that teachers, students, and the general populace alike all understand the power and limitations of these LLM systems as they become more widely adopted.

Ethically, it is the responsibility of LLM developers to offer tools universally and not hide features behind paywalls limiting access to those of lower socioeconomic status. Equality of resources and ease of access to tools that can grant instant access to information and knowledge is paramount. It is also the responsibility of the LLM developers to safeguard against issues of validity, bias, and other ethical issues (Zhuo et al., 2023).

Overall, while the technological advancements of LLMs are promising, it is crucial to address the human factors and ethical considerations involved in their widespread adoption. Improvements in user interface design, education, and accessibility will help bridge the gap between potential and practical use, ensuring that these powerful tools are beneficial and inclusive for all.

The conclusions drawn in this paper highlight the continuous improvement of ChatGPT and other LLMs, drawing a parallel to the advent of search engines. This perspective aligns with findings from other academic papers, such as the comprehensive survey by Guo et al. (2023), which also emphasizes the rapid advancements in LLM capabilities. However, this paper uniquely underscores the gap between the technical generation of outputs and the human ability to create appropriate inputs, a point that is less frequently addressed in other studies.

8.1 Comparison with Existing Research

The observation that most users struggle to effectively interact with generative AI systems is echoed in the work by Chang et al. (2023) and Laskar et al. (2024), which discuss the importance of user interface design in the efficacy of LLMs. These papers focus on the user interface as a critical component of LLM success is a valuable contribution, as it shifts attention to the human-computer interaction aspect, which is often overlooked.

The call for UX specialists to improve LLM interfaces and ensure ethical transactions and accessibility is a significant point of convergence with the broader discourse on AI ethics and

accessibility (Zhuo et al., 2023)). This aligns with the sentiments expressed in previous comprehensive surveys(Guo et al., 2023; Li et al., 2024), which stress the need for rigorous evaluation of LLMs to ensure their safe and beneficial development.

In summary, while the conclusions of this paper resonate with existing literature on the advancements and challenges of LLMs, they also introduce a critical perspective on the importance of user interface design and ethical considerations. This adds a valuable dimension to the ongoing academic conversation about the practical implementation and societal impact of LLMs.

8.2 Practical Applications

The practical implications of this research are substantial, particularly in enhancing the usability and accessibility of large language models (LLMs) like ChatGPT. One of the key findings is the need for improved user interfaces that can bridge the gap between the technical capabilities of LLMs and the average user's ability to generate effective inputs. This research highlights the importance of UX specialists in designing interfaces that are intuitive and user-friendly, ensuring that all users, regardless of their technical expertise, can utilize these powerful tools effectively. The study also underscores the need for comprehensive user education on how to interact with LLMs, which can be integrated into technology curriculums at various educational levels.

The target audience for this research includes UX designers, educators, and developers of LLMs. UX designers can benefit from these insights to create more effective and accessible interfaces, while educators can incorporate findings into teaching materials to better prepare students for interacting with AI systems. LLM developers can use this research to address ethical considerations, ensuring that their tools are accessible to users of all socioeconomic backgrounds and safeguarding against bias and validity issues. By addressing these practical implications, this study contributes to the broader goal of making AI technologies more inclusive and equitable, ultimately benefiting a wide range of users.

8.3 Limitations and Future Research

The experimental study described in this paper had a number of limitations which could be overcome in future research experimentation :

- The participant sample size could be greatly increased to validate the statistical significance of the conclusions found in this study.
- A wider range of participants should be used, in particular targeting those from different backgrounds with diverse speech patterns and languages.
- The study should be expanded to other thematic areas, beyond the realm of Biology, as used in this study.
- Lastly, a comparison of multiple LLMs would be useful – including a comparison of the different releases of ChatGPT.

8.4 Recommendations

Based on the data analysis and conclusions from the experiment described in this paper, a number of UX and interface improvements are proposed :

- Improve the onboarding experience with a user flow that improves understanding of how and what will happen when users interact with the LLM.
- Educate the public on how to appropriately use a range of LLM software.
- Add assistive text algorithms for users to understand the best way to generate input prompts to improve the quality of the output generated.
- Remove paywalls to LLM features to allow for non-biased learning of systems as well as ethical future usage.

9. ACKNOWLEDGEMENTS

The authors would like to thank Dr. April South of the Department of Biological Sciences, University of South Carolina for their assistance. Dr. South allowed her students to engage with ChatGPT as testers and provided invaluable feedback in the grading process of the papers created by students. The authors would also like to thank the students from Dr. South's courses who participated in this study.

10. REFERENCES

Afkarin, M. Y., & Asmara, C. H. (2024). Investigating the implementation of ChatGPT in English language education. Journey: *Journal of English Language and Pedagogy*, 7(1), 57-66.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 15(2).

Baek, T. H., & Kim, M. (2023). Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics*, 83, 102030.

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching*, *6*(2).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45.

Currie, G. M. (2023). Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?. *Seminars in Nuclear Medicine*, 53(5), 719-730.

Dhaini, M., Poelman, W., & Erdogan, E. (2023). Detecting chatgpt: A survey of the state of detecting chatgpt-generated text. *arXiv preprint arXiv:2309*.07689.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301*.04246.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., ... & Xiong, D. (2023). Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:*2310.19736.

Hanses, S., & Wang, J. (2022, April). How do users interact with AI features in the workplace? Understanding the AI feature user journey in enterprise. *In proceedings of 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

Hanna, Robert (2023). How and why ChatGPT failed the Turing Test. *Unpublished MS. Available online atURL*=https://www.academia.edu/94870578/How_and_Why_ChatGPT_ Failed_The_ Turing_Test_January_2023_version.

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and analysis of chat GPT and its impact on different fields of study. *International journal of innovative science and research technology*, 8(3).

Kantorovitch, J., Niskanen, I., Malins, J., Maciver, F.& Didaskalou A. (2017). Supporting The Initial Stages of The Product Design Process: Towards KnowledgeAwareness And Inspiration, International Journal of Human Computer Interaction, 8(1), 8-22.

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, *382*(2270), 20230254.

Kuang, E., Li, M., Fan, M., & Shinohara, K. (2024, May). Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. *In Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J.,& Tseng, V. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health,* 2(2), e0000198.

Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., ... & Huang, J. (2024). A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 13785–13816

Lechien, J. R., Maniaci, A., Gengler, I., Hans, S., Chiesa-Estomba, C. M., & Vaira, L. A. (2024). Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). *European Archives of Oto-Rhino-Laryngology*, 281(4), 2063-2079.

Lee, B. C., & Chung, J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10), 1906-1914.

Lee, S. A., Welch, J., Wallace, R. J., Cross, D., & Loffi, J. M. (2024). ChatGPT in the Classroom: A Practical Guide for Educators. *Journal of Aviation/Aerospace Education & Research,* 33(3), 1.
Skrabut, S. (2023). *80 Ways to use ChatGPT in the classroom: Using AI to enhance teaching and learning*, Skrabut.

Lu, Y., Yang, Y., Zhao, Q., Zhang, C., & Li, T. J. J. (2024). AI Assistance for UX: A Literature Review Through Human-Centered AI. *arXiv preprint arXiv:*2402.06089.

Liu, Y., Xu, Y., & Song, R. (2023). Transforming User Experience (UX) through Artificial Intelligence (AI) in interactive media design, *In proceedings of the IConnect 2023: ES3Web of Conferences* (pp. 1-9).

Martinelli, S., Lopes, L., & Zaina, L. (2024). UX research practices related to long-term UX: A systematic literature review. *Information and Software Technology*, 107431.

Mortazavi, E., Doyon-Poulin, P., Imbeau, D., Taraghi, M., & Robert, J. M. (2024). Exploring the landscape of UX subjective evaluation tools and UX dimensions: A Systematic Literature Review (2010–2021). *Interacting with Computers*, iwae017.

Noever, D., & Ciolino, M. (2022). The turing deception. *arXiv preprint arXiv:2212*.06721.

OpenAI. (2023) *GPT-4 Technical Report*. 2023, ArXiv https://doi.org/10.48550/arXiv.2303.08774

Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in the age of AI: an analysis of ai practitioners' awareness and challenges. *ACM Transactions on Software Engineering and Methodology*, 33(3), 1-35.

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: a comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google bard. *AI*, *4*(4), 949-969.

Rahimi, F., & Abadi, A. T. B. (2023). ChatGPT and publication ethics. *Archives of Medical Research*, 54(3), 272-274.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of applied learning and teaching*, *6*(1), 342-363.

Sahib, T. M., Alyasiri, O. M., Younis, H. A., Akhtom, D., Hayder, I. M., Salisu, S., & Besse, D. M. (2023). A comparison between ChatGPT-3.5 and ChatGPT-4.0 as a tool for paraphrasing English Paragraphs. *In proceedings ofInternational Applied Social Sciences (C-IASOS-2023) Congress* (pp. 471-480).

Shah, A., Iftikhar, S. & Chaudhry, N. (2022). Reviewing Assistive Human-Robot Experiences for Inclusive Human-Robot Interaction, International Journal of Recent Trends in Human Computer Interaction (IJHCI), 11(1), 1-18.

Stige, Å., Zamani, E. D., Mikalef, P., & Zhu, Y. (2024). Artificial intelligence (AI) for user experience (UX) design: a systematic literature review and future research agenda. *Information Technology & People*, 37(6), 2324-2352.

Urban, M., Děchtěrenko, F., Lukavský, J., Hrabalová, V., Svacha, F., Brom, C., & Urban, K. (2024). ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*, 215, 105031.

Vetter, M. A., Lucia, B., Jiang, J., & Othman, M. (2024). Towards a framework for local Interrogation of AI ethics: A case study on text generators, academic integrity, and composing with ChatGPT. *Computers and Composition*, 71, 102831.

Xu, W., Dainoff, M., Ge, L.& Gao, Z. (2023) Transitioning toHuman Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals toEnable Human-Centered AI, *International Journal of Human–Computer Interaction*, 39(3), 494-518,

Xu, Y., Liu, Y., Xu, H., & Tan, H. (2024). AI-driven UX/UI design: Empirical research and applications in FinTech. *International Journal of Innovative Research in Computer Science & Technology*, 12(4), 99-109.

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020, April). Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. *In proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Yang B., Wei L., Pu, Z.(2020). Measuring and Improving User Experience Through Artificial Intelligence-Aided Design. *Frontiers in Psychology*, 11.

Yilmaz, F. G. K., Yilmaz, R., & Ceylan, M. (2024). Generative artificial intelligence acceptance scale: A validity and reliability study. *International Journal of Human–Computer Interaction*, 40(24), 8703-8715.

Zhuo, Y., Li, Z., & Li, H. (2023). Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint arXiv*:2310.19736.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301*.12867, 10(4).