Application of Emerging Technologies, Vol. 115, 2023, 82–93
https://doi.org/10.54941/ahfe1004305

Exploring Generative AI as a Proxy User for Early-Stage User Research – Preliminary Findings

Michael P. Jenkins[1], K. Elizabeth Thiry[2], Richard Stone[3], Caroline Kingsley[4], and Calvin Leather[1]

[1]Knowmadics Inc., Herndon, VA 20171, USA
[2]Self, London, Greater London, NW3 6BB, United Kingdom
[3]Stone Solutions + Research Collective, Scottsdale, AZ 85254, USA
[4]Overmatch Inc., Plano, TX 75024, USA

ABSTRACT

The potential of generative AI has exploded as of late, largely due in part to the improved accessibility that tools like ChatGPT afford non-data scientists and developers. One potential area of application is for generative AI models to serve as proxy users in early-stage user research. User research is a crucial component of product development, helping to understand user needs, preferences, and behaviors. However, conducting user research can be time-consuming and resource-intensive and may require access to a user population that is difficult to access (e.g., military users). Generative AI models have shown remarkable progress in generating human-like text and simulating user interactions based on a significant corpus of training materials that serves as the knowledge base for the AI's reasoning. This paper provides preliminary findings from explorations on the feasibility of leveraging generative AI as a proxy user to inform early-stage user research. Using the GPT-4.0 architecture and the Open-AI ChatGPT user interface (chat.openai.com), we conducted preliminary research for six different candidate end user populations and their respective product concepts. This was accomplished by generating generic product descriptions and notional user personas for each respective product, contextualizing ChatGPT to act as the user persona, and then asking a series of generic user experience research (UXR) questions of the GPT model. Responses from ChatGPT were then scored by three UXR and Human Factors subject-matter experts to evaluate the perceived utility of ChatGPT's responses in terms of supporting early-stage product design as a proxy human user. By evaluating the effectiveness of generative AI as a proxy user, this research aims to shed light on its potential benefits and limitations in supporting early-stage user research efforts. While additional research is still needed (e.g., comparing the results of ChatGPT to responses generated by actual end users and having SME's evaluate the accuracy and completeness of ChatGPT's responses), preliminary findings are promising. Generative AI models hold the potential to serve as a valuable proxy to inform early-stage product design efforts, especially in domains where significant corpuses of data already exist for model training and where access to human end users may be restricted or prohibited.

Keywords: Generative AI, Artificial intelligence, ChatGPT, User experience research, UXR

INTRODUCTION

As technology continues to advance at a rapid pace, the process of designing new products has become more complex and challenging. In this dynamic landscape, understanding user needs and preferences continues to serve as a critical factor for success. Preliminary user research, conducted at the early stages of product development, plays a pivotal role in informing design decisions, enhancing usability, and increasing user satisfaction.

The field of technology design often requires a holistic approach that emphasizes user-centered design. Designers recognize that creating products that resonate with users and meet their specific needs is essential for achieving market success and gaining a competitive edge. By conducting preliminary user research, designers can gain valuable insights into user behaviors, preferences, and pain points, enabling them to develop products that truly address user requirements. This preliminary user research helps to familiarize designers with the users, their expectations, and the environment or domain where candidate technology will be applied. This information serves as a foundation for making informed design decisions, such as feature prioritization, interface design, and overall product architecture. By involving users from the outset, designers can avoid costly redesigns and iterations later in the development cycle, leading to more efficient and cost-effective product development.

Accessing user populations for this contextualizing user experience research can pose several challenges (e.g., recruitment of representative users, monetary or other incentives to compensate users for their time, time constraints associated with product development sprints and/or users' availability, geographic limitations, potential biases or homogeneity in recruited users based on how or where they are sourced, need to establish trust between the user and researcher to obtain genuine responses). Overcoming these challenges requires careful planning, resource allocation, and creativity in recruitment strategies. Employing a combination of methods, such as leveraging user panels, partnering with relevant communities or organizations, and utilizing online platforms for remote research, can help mitigate some of the difficulties associated with accessing user populations for user experience research. However, this is often a resource-intensive process that limits what can be effectively accomplished by organizations operating with real-world financial, staff, and temporal constraints.

Generative AI refers to a class of artificial intelligence (AI) models that can generate new content, such as text, images, or audio, based on the patterns and knowledge learned from training data. These models learn to understand the underlying structure and characteristics of the data and then generate new instances that resemble the training examples. This research was conducted using OpenAI's ChatGPT generative language AI to explore the feasibility of serving as a synthetic user for conducting preliminary user research discovery tasks. This was accomplished by generating generic product descriptions and notional user personas for each respective product, contextualizing ChatGPT to act as the user persona, and then asking a series of generic user experience research (UXR) questions of the GPT model. Responses from ChatGPT were then scored by three UXR and Human Factors subject-matter experts to evaluate the perceived utility of ChatGPT's responses in terms of supporting early-stage product design as a proxy human user.

METHODS

Overview

Our team was comprised of four human factors subject-matter expert (SME) researchers, one of whom is also a certified AI prompt engineer, and a data scientist specializing in AI/ML. Our human factors SMEs have between 7 and 20 years of user experience research (UXR) and/or human factors research experience. To help qualify their responses, three basic questions were asked of each SME providing evaluations to help characterize their propensity to trust and/or adopt new technologies (Table 1).

Table 1. Evaluating UXR SME responses to characterizing questions.

Question	Rater 0	Rater 1	Rater 2	Mean
On a scale of 1 (strongly disagree) to 7 (Strongly Agree), you believe it is acceptable to trust new technologies.	3	3	4	3.33
On a scale of 1 (strongly disagree) to 7 (Strongly Agree), you often try to be an early adopter of emerging technologies.	5	5	6	5.33
On a scale of 1 (strongly disagree) to 7 (Strongly Agree), A.I., and LLMs in particular, can be trusted to provide accurate and reliable responses if trained on a sufficient corpus of relevant data.	3	4	5	4.00

Our high-level approach for this effort involved conducting UXR interviews for six respective products and services. Interviews were conducted using ChatGPT as a proxy end-user with the goal of informing the design and/or requirements for the notional products and services as if conducting traditional UXR with representative end-user participants or volunteers during early product design efforts.

The goals of this effort were to: (i) conduct an initial evaluation to determine if the information that ChatGPT provided was valuable for early stage UXR; (ii) determine if there were areas where ChatGPT was particularly useful or lacking; and (iii) establish an initial corpus of data that can be used for a follow-on study involving human subject-matter expert participants to directly compare AI to humans in terms of the benefits AI may afford for UXR.

To accomplish these research goals, our method involved the following steps.

1. Developed six distinct conceptual products and services to serve as the basis for contextualizing the UXR, each designed to explore a unique domain or UXR challenge to enable evaluation of how ChatGPT performs.
2. Our UXR experts rated those products on several factors based on how challenging they perceived the product concept to be to conduct preliminary UXR.

3. Generated generic UXR questions that would be appropriate for interviewing a human participant to elicit insights that inform the design of the respective new products and services. (Note that for consistency across products, no follow-up questions were allowed, which is a known limitation of this research and its comparison to traditional UXR methods).
4. Generated a detailed User Persona to represent a user of each candidate product or service.
5. Prompted ChatGPT to act as the representative user by training it on the User Persona.
6. One team member conducted the UXR interview with ChatGPT (GPT-4.0).
7. The three other UXR and Human Factors SMEs on the team scored responses from ChatGPT using a standardized scoring approach (detailed below) to determine the utility of the AI's responses to inform product design.

ChatGPT Generative AI

ChatGPT was selected as our generative AI for this research effort due to its current popularity and the wide range of knowledge it has been trained on. ChatGPT is powered by OpenAI's GPT (Generative Pre-trained Transformer) model, GPT-4.0, their latest state-of-the-art language model. It is designed to generate human-like text responses based on the input it receives. During training, the model is exposed to a vast amount of text data from the internet, including books, articles, websites, and more. By learning patterns and relationships within the text, the model develops an understanding of language, grammar, context, and even some level of common-sense reasoning. When users interact with ChatGPT, they provide a text-based prompt or message as input. The model then processes that input and generates a response based on its training and learned knowledge. The response is not predetermined or scripted but is generated dynamically by the model. ChatGPT uses a combination of pattern recognition, statistical analysis, and predictive modeling to generate coherent and contextually appropriate responses.

Conceptual Products

Table 2 provides a description of the conceptual products that were used for the UXR interviews. The table also denotes the target user population and justification for inclusion in this research effort, respectively, for each conceptual product. The fourth column of the table also includes the mean scores from the SMEs that were used to rate the UXR challenge for the respective products, i.e., how challenging it would be to conduct traditional preliminary UXR given the nature of the product and its target user population. These scores were generated by asking for ratings from 1 (strongly disagree) to 7 (strongly agree), given four distinct statements (provided below) that were meant to characterize different aspects of UXR difficulty (i.e., higher scores equate to more challenging UXR research). Through this characterization, we were able to establish that the six different notional products covered a range (with mean scores of 2.75 through 5.50) of UXR difficulties.

Table 2. Notional products conceptualized to contextualize the UXR.

ID	Product Description	Inclusion Justification	SME Ratings
1	Point of Sale (POS) software application for a restaurant or bar intended to allow waitstaff to easily: (1) input customer orders so they can be sent to the kitchen or bar staff for order creation, and (2) generate bills for customers.	*Class:* Client software application *HMI:* Touchscreen *Language:* Common / Non-Technical *AI Training Content:* Saturated Product Market, Lacking Procedural Documents, Minimal Research Publications *Users:* Waitstaff and Bartenders *Persona:* "Sarah Thompson"	**Overall:** 2.75 **User Access:** 2.00 **Resource Cost:** 1.33 **Time:** 2.67 ***n* of 3:** 5.00
2	Internet of Things (IoT) wearable for military JTAC users that enables hand gestures to be performed to control a set of predefined functions within the Android Team Awareness Kit (ATAK) application.	*Class:* Embedded software + hardware *HMI:* Gesture *Language:* Technical, Domain Nuanced *AI Training Content:* Analogous Commercial Products, Procedural Documents, ATAK Documentation, Salient HMI Research Publications *Users:* Joint Terminal Attack Controllers (JTACs) in the US Air Force *Persona:* "Capt. Mark Reynolds"	**Overall:** 5.50 **User Access:** 5.67 **Resource Cost:** 6.00 **Time:** 6.00 ***n* of 3:** 4.33
3	Smartphone or tablet application that can track an individual's movement during Pilates to provide insights for instructors to easily: (1) create personalized and tailored regimens for each individual client, (2) track form, movement patterns and progress, and (3) integrate IoT devices for potential virtual personalized instruction or connected exercise equipment.	*Class:* Smartphone / tablet software *HMI:* Touchscreen *Language:* Common, Domain Nuanced *AI Training Content:* Skewed domain documentation (i.e., details on Pilates, but minimal documentation on being a good instructor for Pilates), procedural steps that describe a physical task, widespread domain materials (exercise and fitness), minimal academic research. *Users:* Pilates instructors *Persona:* "Emily Rodriguez"	**Overall:** 2.83 **User Access:** 2.67 **Resource Cost:** 2.00 **Time:** 2.67 ***n* of 3:** 4.00
4	AI-powered smart medicine scheduling and dosing application to aid nurses with in-patient care, which also integrates with patient health monitoring systems as source of data for the AI.	*Class:* Software *HMI:* Keyboard/Mouse *Language:* Technical, Domain Nuanced *AI Training Content:* Highly technical domain knowledge, volumes of training and academic research, analogous product data. *Users:* Nurses serving in hospitals for in-patient care *Persona:* "Martin Rodriguez"	**Overall:** 4.75 **User Access:** 5.00 **Resource Cost:** 4.33 **Time:** 5.67 ***n* of 3:** 4.00
5	A torque wrench with a digital torque sensor intended for hobbyist mechanics that care for their own vehicles.	*Class:* Primarily Hardware *HMI:* None *Language:* Common, Domain Nuanced *AI Training Content:* Usage / application details, procedural relevance to wide range of tasks, widespread domain materials (DIY mechanic) *Users:* Hobbyist / DIY mechanics *Persona:* "Emily Rodriguez"	**Overall:** 3.42 **User Access:** 4.00 **Resource Cost:** 4.00 **Time:** 3.33 ***n* of 3:** 2.33
6	An AI-based wearable personal assistant for Japanese adults that is intended to be used outside the home.	*Class:* Hardware & Software *HMI:* Speech *Language:* Non-English (Japanese), Culturally Nuanced *AI Training Content:* Large volumes of general information on Japanese culture, Translation volumes, Analogous technology product information, minimal academic research, very broad application space. *Users:* Japanese adults *Persona:* "Hiroshi Sato"	**Overall:** 5.00 **User Access:** 4.33 **Resource Cost:** 4.67 **Time:** 5.67 ***n* of 3:** 4.04

1. **User Access:** For the given product, indicate the degree to which you believe it would be ***challenging to access representative users*** to conduct preliminary user experience research for the selected product.
2. **Resource Cost:** For the given product, indicate the degree to which you believe it would be ***costly to recruit and schedule representative users*** to conduct preliminary user experience research for the selected product.
3. **Time:** For the given product, indicate the degree to which you believe it would be ***time-consuming to conduct UXR interviews*** with representative users for the selected product.
4. ***n* of 3:** For the given product, indicate the degree to which you believe ***speaking with 3 or fewer users would be sufficient*** to conduct preliminary user experience research for the selected product.

User Personas

To prime ChatGPT to better serve as a proxy end user, ChatGPT was instructed to act as a user based on a detailed user persona (see Table 3 for one example). This approach enables the AI to provide a unique filter on the knowledge it will put to use and gives a voice and perspective to the AI's responses that enable it to better play the role of a proxy user during UXR.

Table 3. One of the six user personas that were used to prime ChatGPT to serve as a proxy end user (the remaining five were omitted due to page length constraints).

Persona Details
Product ID Target: 1
Name: Sarah Thompson
Age: 28
Gender: Female
Profession: Waitress in a busy city restaurant
Location: Currently living and working in New York City, New York, but originally from a small town in Vermont.
Family: Sarah is single and has no children. She is the eldest of three siblings and moved to New York City to explore job opportunities and the diverse culinary scene. Her parents still reside in Vermont.
Income Level: Sarah's income primarily depends on the hourly wage for waitstaff and tips from customers. Despite the variability in her income due to the nature of her job, she manages to lead a comfortable lifestyle within the bustling city.
Education: Sarah completed her high school education in Vermont. While she hasn't pursued a formal college degree, she has taken several culinary and hospitality courses online and in-person to increase her knowledge and skills in the food industry.
Tech Savvy: Sarah is competent in using technology necessary for her job, such as digital POS (Point of Sale) systems, and tablets for taking orders. She also uses her smartphone and laptop for personal purposes like social media, staying informed about industry trends, and online learning.
Languages: Sarah is fluent in English. Given the diversity of the New York City, she has learned basic Spanish phrases to assist a broader range of customers.
Background: Sarah grew up in a small town and moved to the city after completing high school. She has been working as a waiter for the past six years, gaining experience in different restaurant settings. She is passionate about food and has developed a deep understanding of the restaurant industry.

Table 3. Continued.

Goals and Motivations:

1. Providing excellent customer service: Sarah takes pride in delivering exceptional dining experiences to customers. She strives to ensure their satisfaction by actively listening to their needs and going the extra mile to exceed their expectations.
2. Enhancing teamwork and collaboration: Sarah enjoys working with a diverse team of restaurant staff. She believes that effective teamwork is essential for delivering high-quality service and creating a positive work environment.
3. Continuous learning and improvement: Sarah is always seeking ways to enhance her skills as a waiter. She stays updated on the latest menu offerings, learns about food and wine pairings, and explores new techniques to enhance the dining experience.

Challenges:

1. Time management: Working in a busy restaurant chain requires Sarah to handle multiple tables simultaneously while ensuring prompt service. Managing time efficiently can be challenging, especially during peak hours.
2. Dealing with difficult customers: Occasionally, Sarah encounters demanding or irate customers. She must remain composed and find ways to resolve conflicts while maintaining a positive attitude.
3. Staying up to date with changes: Popular restaurant chains often introduce new menu items, policies, or technology. Sarah needs to adapt quickly and stay informed to provide accurate information and offer the best service to customers.

Skills and Abilities:

1. Strong interpersonal skills: Sarah is friendly, approachable, and able to connect with customers on a personal level. She communicates clearly and effectively, making customers feel comfortable and valued.
2. Attention to detail: Sarah pays close attention to customers' orders, dietary restrictions, and special requests. She ensures accuracy in food preparation and presentation, minimizing errors or inconsistencies.
3. Problem-solving: Sarah can think on her feet and handle unexpected situations. Whether it's a mix-up in orders or an unhappy customer, she remains calm, seeks solutions, and involves the necessary team members when needed.

Lifestyle: Sarah lives a fairly busy lifestyle due to the high-paced nature of her job. When she is not working, she enjoys exploring New York's vibrant culinary scene, attending food festivals, and trying out new recipes at home. She also enjoys yoga and jogging in the park to maintain her health and manage stress.

Current Tech Usage: At work, she uses digital systems to manage orders and transactions. Personally, she relies heavily on her smartphone for various activities, such as social media, email, online shopping, streaming music and movies, and staying updated with online food and beverage publications. She also uses productivity and mindfulness apps to manage her schedule and stress levels.

Key Quote: "I love creating memorable dining experiences for our guests. It's rewarding to see their smiles and know that I contributed to their enjoyment. Every day brings new challenges and opportunities to grow as a waiter."

UXR Questions

Development of the UXR questions started by generating a set of generic research questions to guide interviews with users. These questions served as exemplars for what UXR researchers would use as a starting point to explore user perspectives, preferences, and pain points when informing a new product design. To extract useful and relevant information, each question was modified with placeholders where the job that the notional product sought to aid was swapped in, and likewise the product that was being designed. Table 4 provides the UXR questionnaire. While not an exhaustive list, this question set was designed to ask questions that covered three key areas that all SMEs on our team agreed were core to UXR: (1) contextualizing how the job or work is done today that the product seeks to aid/augment; (2) drawing on specific experiences of the user, good and bad, where the product would provide value; and (3) characterizing how the product would integrate into current tasks and workflows. To limit the complexity of this preliminary investigation, we did not allow the interviewer to ask follow-up questions (e.g., for clarification or to explore specific topics or information) as would normally be done with a UXR interview.

Table 4. UXR questionnaire.

ID	UXR Question
1	Can you describe a typical day or scenario in the role of a [**JOB**] where you believe they would potentially use a [**PRODUCT**]?
2	What technology does [**JOB**] currently rely upon that would be replaced or augmented by a [**PRODUCT**]?
3	What similar products or solutions do [**JOB**] use that offer similar capabilities or features to a [**PRODUCT**]?
4	What works well and doesn't work well about these alternative solutions?
5	Try and recall a time or event when a [**JOB**] had a particularly positive experience with a similar product. Can you describe that time and what made it a positive experience?
6	What challenges or frustrations do [**JOB**] currently face when trying to achieve their goals with those current products or solutions?
7	Can you envision any new capabilities or features for a [**PRODUCT**] that would enhance the experience for [**JOB**] or make their tasks easier?
8	What are performance measures that will make a [**JOB**] want to use a [**PRODUCT**] like this? And how would you prioritize those measures?
9	How frequently would you expect [**JOB**] to use a [**PRODUCT**] in their daily or regular activities?
10	How do you envision [**JOB**] integrating a [**PRODUCT**] into their current workflow, equipment, and lifestyle?
11	Can you provide a few very specific examples of scenarios, tasks, or actions where a [**PRODUCT**] would be particularly valuable to [**JOB**]**?**
12	Can you provide a few very specific examples of working environment contexts under which a [**PRODUCT**] would be particularly valuable to [**JOB**]?
13	Can you think of any potential drawbacks or concerns that might arise from [**JOB**] incorporating a [**PRODUCT**]?
14	Is there anything else you would like to share about [**JOB**] needs, preferences, or experiences related to this novel a [**PRODUCT**]?

Data Collection

Three of our team's UXR SMEs independently rated each ChatGPT response to the 14 questions posed for each of the six products (252 total ratings). Ratings were based on a Likert scale from 1 (Strongly Disagree) to 7 (Strongly Agree) given the prompt: "*Indicate the degree to which you believe the information provided in response to the given question provides significant utility* *to inform initial requirements, features, and/or designs for the novel product or service offering presented*".

RESULTS

Figure 1 (average per product) and Figure 2 (by rater and individual question) show the rating results. Overall, the mean rating given was 5.69, indicating that raters agreed that ChatGPT responses provided significant utility to inform initial product design tasks across all six product domains. As Figure 1 illustrates, even within the individual product domains/categories, the mean rating was always > 5, suggesting that even across these disparate product domains, ChatGPT responses provided UXR utility.

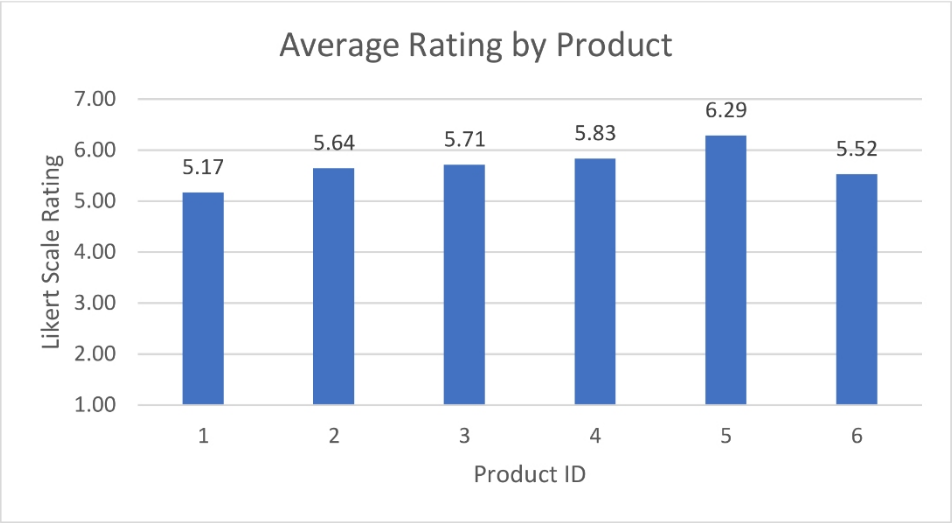

Figure 1: Average rating on the 1 (strongly disagree) to 7 (strongly agree) indicating the degree to which raters felt ChatGPT responses to given questions provided "*significant utility to inform initial requirements, features, and/or designs*" of the respective products.

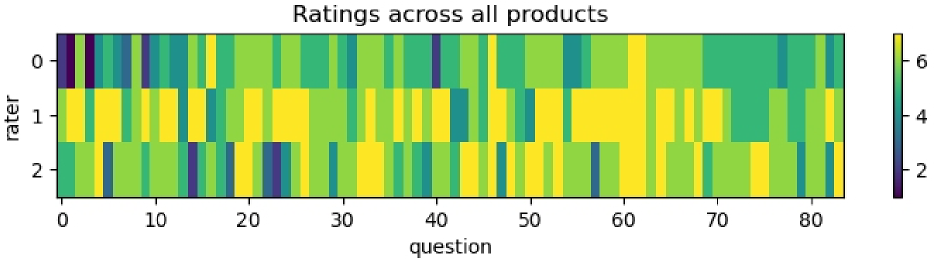

Figure 2: Individual ratings by rater across all products for each of the 84 questions asked (i.e., questions 0-13 = Product 1).

Figure 2 also shows that rater 0 more strongly disagreed with ChatGPT responses for Product 1 (i.e., questions 0–13 in Figure 2). Upon further investigation into the rater's response justifications, this was because the rater had significant prior experience working in the service industry (as wait staff). While excluding these responses for Product 1 increases the mean rating for that product to 5.93, the rater summarized the lower scores as a result of ChatGPT's responses being too generic, suggesting a lot of features that the majority of POS systems already incorporate, and failing to identify real pain points the novel product should address. Similarly, for Product 2 (i.e., questions 14–27 in Figure 2), rater 3 had direct working experience in the product domain and thus provided objectively lower scores than the other two raters (i.e., removing the rater's responses for Product 2 increases the mean score to 6.00). The rater's explanation for these scores is: "*This is a solid entry point for a very niche field. It can serve as a starting point until an end user is found. The errors are small enough that the majority of military members would not catch it, but as a [qualified and current] JTAC, I can see how this is glossing over key points*". While this is limited data to draw conclusions, there is a potential correlation/implication for further investigation to characterize the diminishing returns beyond ChatGPT response that adding a representative user to UXR would yield. In the two cases presented here, the domain-qualified raters resulted in the perceived utility of ChatGPT responses decreasing by 16.0%. The other perspective that may be appropriate is that AI and LLMs provide information that may be misleading for UXR researchers who lack explicit domain experience to verify the appropriateness and robustness of AI responses. Again, implying the need to include a domain SME as part of the UXR process, through an interview or potentially just to fact-check AI responses, has potential value.

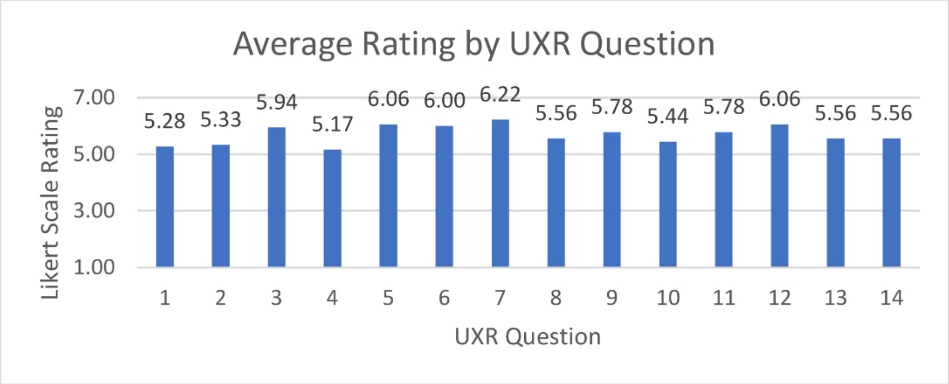

Figure 3: Average rating on the 1 (strongly disagree) to 7 (strongly agree) indicating the degree to which raters felt ChatGPT responses to given questions provided "*significant utility to inform initial requirements, features, and/or designs*" for each UXR question.

Figure 3 provides average ratings for each of the 14 UXR questions that were asked. Among the questions, the lowest score was given to question 4, which asked what worked well and what did not work well about similar products or solutions that were already in use (i.e., incumbent or competitor products to the one being proposed). While subject to further investigation, this may be due to the fact that question 4 was the only question (by design) that was included as a progressive prompt (i.e., question 4 was designed to prompt the AI to recall aspects of the response it provided to question 3 to contextualize its response to question 4).

For questions that were scored above a 6.00 (i.e., questions 5, 6, 7, and 12), it is notable that these questions are largely focused on prompting the AI to provide specific experiences or examples of when, how, or why the product would or would not be useful. While the AI clearly has never experienced any of the work domains or alternative products firsthand, the responses that were deemed most valuable were those where it generated synthetic examples of these experiences. This contrasts with other questions that focused on eliciting more rote information (e.g., useful features of competitive, alternative, or current products).

Finally, to assess the agreement between the raters in categorizing responses into the Likert scale categories, we first used Fleiss' kappa, which yielded a score of 0.011. Since the range of Fleiss' Kappa is −1 to 1, a score of 0.011 indicates that the observed agreement among the raters is only slightly better than expected due to random chance. This suggests that there is not much consistency or agreement in the way the raters are categorizing the utility of ChatGPT responses. Similarly, the slightly more appropriate intraclass correlation coefficient (ICC(1,3)) resulted in a score of 0.187674, indicating the agreement among the raters can only be considered slight. This means that there was relatively low consistency or agreement among the assessed utility provided by the three raters. However, interpretation of this finding is key to understanding the implications for future research. These low scores across the questions indicate that there was little correlation beyond chance between what responses raters viewed as useful versus less useful. Therefore, while the three raters all found the responses overall highly useful (i.e., the 5.69 overall mean rating), different raters saw different questions as providing more utility than others, with correlation only slightly better than chance.

CONCLUSION

This work represents only a first step to evaluating the feasibility of using LLMs and other generative AI models to aid in UXR. The results are generally promising, indicating that some potential exists for AI and LLMs to aid in UXR; however, defining where that value lies and how to best extract it still requires additional research. Moving forward, we believe that the next steps in our research will include: (1) securing IRB approval to have human SMEs for given product areas respond to the same questions for comparative analysis; (2) having the qualified SMEs and domain experts evaluate the responses of AI outputs for accuracy and robustness; (3) enabling follow-up questions that would be appropriate for normal UXR; and (4) repeating the methods with an LLM or generative AI model trained exclusively on content specific to a given product's domain of application (to determine if tuning the model improves its performance). These direct comparisons and tuning of the collection methods will afford more precision in the analysis to identify where the value lies and how to access it.

If LLMs and AI can in fact be designed to serve as proxy end users, the potential benefits are significant: AI never fatigues, can be more descriptive in responses (less probing), can play the role of multiple distinct end users, may be more cost-effective, is non-litigious/not subject to potential psychological risks, is nearly always available, and has the potential to accurately represent the opinions of an entire population of opinions upon which it is trained. However, potential drawbacks also need to be taken into consideration when relying on its outputs: can become stale or biased based on training data, may be inaccurate or generate nonsensical results, and has limited memory to recall its prior responses or context of UXR. Currently, we believe it is most important to note that while ChatGPT can generate impressive and often helpful responses for UXR, it can also produce incorrect or nonsensical answers. It lacks true understanding or real-world knowledge and relies solely on the patterns it has learned from the training data. While ChatGPT and similar LLMs and AI may be an attractive option, our results point to the need to validate its responses with a domain expert or representative user until a time when its performance or pedigree can be better established.

ACKNOWLEDGMENT

The authors would like to acknowledge the developers at OpenAI who are responsible for the development of the ChatGPT generative AI service that was used to conduct this research effort.

REFERENCES

Jamieson, S. (2004). Likert scales: How to (ab) use them?. *Medical Education*, *38*(12), 1217–1218.

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in health sciences education*, *15*, 625–632.