How can LLMs support UX Practitioners with image-related tasks?

RUICAN ZHONG, University of Washington, USA
GARY HSIEH, University of Washington, USA
DAVID W. MCDONALD, University of Washington, USA

There has been increasing attention on using LLMs (large language models) to assist text-based UX design processes, such as brainstorming and ideation. Given LLMs' recent ability to process images and summarize image content, they might also be leveraged to conduct heuristic evaluations of UIs (user interfaces). As an initial step, our work tested how well LLMs could conduct synthetic heuristic evaluations. We found that LLMs are able to assess the components of UIs according to a given set of heuristics. We present insights about how LLMs' outputs compare to human insights and hope to discuss how UX designers can leverage LLMs in their current process of usability testing.

1 INTRODUCTION

There has been increasing attention to understanding the use of LLMs in UX design [7, 33]. Prior work that applies LLMs to the work of UX designers has focused on tasks that can be represented in text. As multimodal LLMs become available, there is an opportunity to explore the use of LLMs in image-related tasks, such as heuristic evaluation of UIs (user interfaces).

Recent research has shown that multimodal LLMs can identify and describe specific elements of the UIs [13, 30, 33]. In addition, since LLMs have been trained on a huge amount of information (e.g., scientific papers, discussion forums, etc.), they can possibly model how humans think about a specific problem [31]. Thus, it is possible that multimodal LLMs can take on complex UX tasks related to images, such as heuristic evaluation.

Heuristic evaluation is a form of user evaluation, where potential users are given a user interface. After inspecting and interacting with the interface, they are asked to identify if there were any violations of a set of heuristics. The heuristics are usually an established set of standards that are widely used. And the identified violations are referred to as usability issues [23, 24]. As described, multimodal LLMs could process UIs and identify elements, demonstrating their potential to assess specific components against a given set of content. Additionally, drawing on LLMs' ability to simulate human cognition, they could possibly provide feedback about how each element follows or violates a set of heuristics. Overall, given LLMs' recent abilities to process images, it is beneficial to understand how they could support UX practitioners with image-related tasks.

In this work, we explore the feasibility of LLMs in conducting synthetic heuristic evaluations by testing their performance on two common types of mobile applications. We present findings demonstrating that LLMs could generate qualitative heuristic evaluation results and describe the usability issues, in a similar way as human evaluators. However, we also note that simply understanding the ability of LLMs is not enough. It is important to consider how to utilize LLMs to support designers' work. We discuss possible directions exploring how to integrate LLMs to support UX designers in synthetic heuristic evaluations.

2 RELATED WORKS

2.1 Heuristic Evaluation

Heuristic evaluation is a commonly used formative evaluation technique [23]. To conduct a heuristic evaluation, an evaluator is provided with a list of heuristics and is tasked to review the interface design and note down any issues violating the provided heuristics – a set of principles commonly used to ensure the usability of the interface [24]. One of the most common sets of heuristics is Neilsen's 10 heuristics. It was created by Jakob Neilsen and Molich in the 1990s [21, 23]. For each violation found, evaluators are also asked to rate the severity of the usability issues from 0 to 4 to help prioritize the most important issues to fix.

In our work, we chose heuristic evaluation as the user evaluation technique to automate using LLM for several reasons. First, heuristic evaluation is a flexible technique that has been shown to be effective for evaluating web and mobile interfaces [1, 2, 5, 8, 11]. By adapting the set of 10 original heuristics by Nielsen et al., heuristic evaluation can also be extended to a variety of novel interfaces and interactions [17–19]. Second, heuristic evaluation does not require the evaluators to be users or have experience with the interface evaluated [23]. This prevents the need from having domain specialized LLMs. Third, heuristic evaluation can be done by inspecting screenshots, and does not require interactions with the actual interface [12]. Analyses of screenshots is now possible with GPT-4, while more complex interactions are not yet directly possible and would require additional engineering.

2.2 Automated Usability Testing

According to Ivory & Hearst's 2001 survey paper on automating user evaluation, of the 75 usability evaluation methods surveyed, methods with automation support only accounted for 33% of the methods [15]. The two types of methods where automation support is most prevalent are Analytical Modeling (e.g., GOMS analysis) and Simulation (e.g., information processing analyses). The use of quantitative metrics and behavioral trace logs in these methods inherently supports automated capture and analyses. However, these techniques can be difficult to use and learn as they require the construction and manipulation of complex models. Further, the focus on task completion and usage traces "does not capture important qualitative and subjective information (such as user preferences and misconceptions) that can only be unveiled via usability testing, heuristic evaluation, and other standard inquiry methods" [15].

A more recent review of automated web usability evaluation tools suggested that many of these problems persist [22]. The 10 popular web usability testing tools examined generally only produce a score as output, offering no reflection of the meaning of that output [22]. Further, the high variance in the evaluation results also raised concerns about the consistency and accuracy of the tools, and whether usability is actually assessed versus other more easily quantifiable features, such as performance (e.g., page load speed), SEO, and page size.

LLMs offer a potential solution to help advance this line of research. There has been a growing interest in applying LLMs in UX design processes given LLMs' capabilities [3, 4, 6, 9, 10, 14, 16, 28, 32]. First, LLMs are able to process and analyze images [13, 30, 33]. For instance, Zeng et al. demonstrated that given an image and a short description of the image, multimodal LLMs could provide a summary description of the image content [33]. Additionally, LLMs are keen to provide feedback [7]. Since LLMs have been trained on a large corpus of data, they could draw similar conclusions as human beings and generate human-like ideas [31].

However, as these models evolve over time, their capability may change and the results may change. In addition, compared to the existing research's [7, 13, 30, 33] focus, heuristic evaluation is more of a cognitive task than a straightforward question-and-answer engagement. Thus, based on these existing works, we see a potential for LLMs

to address the gap in automated usability testing and, more specifically, synthetic heuristic evaluations. But we still have open questions to explore. For instance, can LLMs produce consistent and reliable results? Can LLMs avoid overemphasizing some given evaluative dimensions more than others in synthetic evaluations? Can LLMs produce interpretative feedback, similar to that of human evaluators? Our work hypothesizes that the underlying cognitive ability of LLMs, indicated by Schmidt et al. [31], can support their use in synthetic evaluations.

3 PROCEDURE

We selected two mobile applications to test the performance of LLMs in synthetic heuristic evaluations. The first one is a rental app. The second one is a language learning app. For each app, we asked both human evaluators and GPT-4 to experience some tasks, such as *"set up rental search preferences," "search for an apartment using a criteria,"* and *"experience a French learning lesson."* We selected these two apps to cover as many different interactions and UI elements as possible. For each task, we took 3-9 screenshots to demonstrate the screens that users would encounter to complete the task.

To understand how well LLM could conduct heuristic evaluations, we collected two sets of evaluation results. The first one is the synthetic evaluation set, and the second one is the master set. We compared synthetic evlauation outputs against the master set.

3.1 Synthetic evaluation set

We used the screenshots taken to prompt GPT-4 to complete a synthetic evaluation of both the rental app and the language learning app. We chose GPT-4 because it is a state-of-the-art multimodal LLM [25, 26]. We iterated our instructions a few times to ensure that the outputs were consistent and of high quality. Then, we grouped the results by the usability issues identified and analyzed them.

3.2 Master set

The master set of heuristic issues represents the complete list of all heuristic violations of the two apps uncovered in our study.

To gather this set, we collected responses from five local research assistants, who had been trained on conducting heuristic evaluation. We asked them to conduct independent evaluations of the rental app and language learning app. They then came together to discuss the issues they identified and those were added to compile the master set. In the master set, there were a total of 56 issues found for the rental app and 48 issues found for the language learning app.

4 RESULTS

We compared the synthetic evaluation results against the master set. The synthetic evaluation was able to uncover 64.28% (36/56) of the usability issues in the rental app. And the synthetic evaluation detected 62.50% (30/48) of the usability issues in the language learning app.

We also contrasted LLMs' and the master set's description of the same usability issue. For instance, when evaluating the first set of screens of the rental app, the master set mentioned that *"The phrase 'find your for now and forever' is not clear in its meaning because it is abstract and generic, having no explicit association with the rental context."* Our synthetic evaluation similarly pointed out that *"The slogan 'Find your forever. Or your for now.' might be confusing as it's not immediately clear that it refers to the duration of property rentals."* In general, we found that the LLM was able to provide a human-like description of the usability issues identified.

5 DISCUSSION

In this workshop, we hope to engage in conversations about the feasibility of using generative AI to conduct synthetic heuristic evaluations, the role of LLMs in performing UX design tasks, and how to measure LLMs' performances given their uncertain nature.

5.1 Cognitive Capability of LLMs

The initial testing of LLMs' performance in conducting synthetic heuristic evaluation shows their capability to assess user interface components and corroborates LLMs' potential to simulate cognitive capabilities of humans [31]. In comparison to existing automated evaluation systems [15, 22], LLMs were able to provide interpretable qualitative insights similar to human feedback, which addresses some of the difficulties with prior automated usability testing systems [15, 22] that only produced a score with no explanation.

To further understand LLMs' cognitive capacity, more work could be done to contrast LLMs' results and human outputs. Specifically, we could collect human expert evaluators' assessment of user interfaces and directly contrast it with LLMs' outputs. Comparing the qualitative feedback from humans and LLMs would help us better understand whether LLMs can "think" similarly to human beings. This may help us identify which aspect of synthetic evaluations LLMs fails to address. For instance, LLMs may have difficulty understanding a specific heuristic and do not perform as well as human beings in that category. This comparison may also tease out some advantages of using LLMs to do heuristic evaluations. Since LLMs do not suffer from fatigue and repetitive work, they may be able to produce high-quality results more consistently than humans.

5.2 Integrating LLMs in Synthetic Heuristic Evaluation

Since LLMs can feasibly run a heuristic evaluation by themselves, designers may ask LLMs to conduct initial evaluations, test, and iterate the interfaces before moving forward to official testing with human evaluators. This could increase the quality of human testing sessions, as the iteration with LLMs would have at least identified the obvious set of issues. As we further fine-tune and improve the quality of LLMs' outputs, replacing human labor might also be possible. But more importantly, it is valuable to explore the role of LLMs in the design process. How would leveraging LLMs' insights for heuristic evaluations change the way designers brainstorm and iterate their products? How would using LLMs change the collaboration dynamics in teams of UX designers?

5.3 Measuring Reliability

More work is also needed to ensure LLMs can produce consistent and reliable results. As prior works have pointed out [20, 27, 29], LLMs are inherently stochastic and do not output the same result every time. But for tasks like heuristic evaluation whose goal is to identify the same set of usability issues, it may not matter if the results are exactly the same. As long as LLMs can consistently identify the same set of usability, the phrasings could be different and would not impact the evaluation result. This opens up the question of how to ensure the consistency of LLM outputs. How can we prompt LLMs so that they would identify a consistent set of issues given the same user interfaces? Furthermore, how should we measure LLMs' reliability in doing synthetic evaluation? More work is needed to systematically evaluate the use of LLMs in UI evaluations. We are interested in having these discussions at the workshop.

REFERENCES

[1] Sai Pankaj Akula. 2021. A critical evaluation on SRK STORE APP by using the Heuristic Principles of Usability.

[2] Hend S Al-Khalifa, Bashayer Al-Twaim, and Bedour AlHarbi. 2016. A heuristic checklist for usability evaluation of Saudi government mobile applications. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*. 375–378.

[3] Salvatore Andolina, Khalil Klouche, Diogo Cabral, Tuukka Ruotsalo, and Giulio Jacucci. 2015. InspirationWall: supporting idea generation through automatic information exploration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*. 103–106.

[4] Salvatore Andolina, Hendrik Schneider, Joel Chan, Khalil Klouche, Giulio Jacucci, and Steven Dow. 2017. Crowdboard: augmenting in-person idea generation with real-time crowds. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 106–118.

[5] Ahmad Azizi, Mahmood Maniati, Hadis Ghanbari-Adivi, Zeinab Aghajari, Sedigheh Hashemi, Bahareh Hajipoor, Asma Rabiee Qolami, Maryam Qolami, and Amirabbas Azizi. 2021. Usability evaluation of hospital information system according to heuristic evaluation. *Frontiers in Health Informatics* 10, 1 (2021), 69.

[6] Suyun Sandra Bae, Oh-Hyun Kwon, Senthil Chandrasegaran, and Kwan-Liu Ma. 2020. Spinneret: Aiding creative ideation through non-obvious concept associations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[7] Michael Balas, Jordan Joseph Wadden, Philip C Hébert, Eric Mathison, Marika D Warren, Victoria Seavilleklein, Daniel Wyzynski, Alison Callahan, Sean A Crawford, Parnian Arjmand, et al. 2023. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *Journal of Medical Ethics* (2023).

[8] André Castello Branco, Eveline Sacramento, Eliza Oliveira, Oksana Tymoshchuk, Maria Antunes, Margarida Almeida, Luís Pedro, Fernando Ramos, and Daniel Carvalho. 2022. Usability Evaluation of a Community-led Innovation Mobile App. (2022).

[9] Sara Bunian, Kai Li, Chaima Jemmali, Casper Harteveld, Yun Fu, and Magy Seif Seif El-Nasr. 2021. Vins: Visual search for mobile user interface design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[10] Graham Dove and Sara Jones. 2014. Using data to stimulate creative thinking in the design of new products and services. In *Proceedings of the 2014 conference on Designing interactive systems*. 443–452.

[11] Reese Hoi Yin Fung, Dickson KW Chiu, Eddie HT Ko, Kevin KW Ho, and Patrick Lo. 2016. Heuristic usability evaluation of university of Hong Kong libraries' mobile website. *The Journal of Academic Librarianship* 42, 5 (2016), 581–594.

[12] Nielsen Normal Group. 2023. How to Conduct a Heuristic Evaluation. https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/#:~:text=A%20heuristic%20evaluation%20is%20a,make%20systems%20easy%20to%20use.

[13] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[14] Angel Hsing-Chi Hwang. 2022. Too late to be creative? AI-empowered tools in creative processes. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–9.

[15] Melody Y Ivory and Marti A Hearst. 2001. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)* 33, 4 (2001), 470–516.

[16] Peter Kun, Ingrid Mulder, Amalia De Götzen, and Gerd Kortuem. 2019. Creative data work in the design process. In *Proceedings of the 2019 on Creativity and Cognition*. 346–358.

[17] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[18] George D Magoulas, Sherry Y Chen, and Kyparissia A Papanikolaou. 2003. Integrating layered and heuristic evaluation for adaptive learning environments. In *Proceedings of the second workshop on empirical evaluation of adaptive systems, held at the 9th international conference on user modeling UM2003, Pittsburgh*. 5–14.

[19] Jennifer Mankoff, Anind K Dey, Gary Hsieh, Julie Kientz, Scott Lederer, and Morgan Ames. 2003. Heuristic evaluation of ambient displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 169–176.

[20] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. *arXiv preprint arXiv:2305.14552* (2023).

[21] Rolf Molich and Jakob Nielsen. 1990. Improving a human-computer dialogue. *Commun. ACM* 33, 3 (1990), 338–348.

[22] Abdallah Namoun, Ahmed Alrehaili, and Ali Tufail. 2021. A Review of Automated Website Usability Evaluation Tools: Research Issues and Challenges. In *International Conference on Human-Computer Interaction*. Springer, 292–311.

[23] Jakob Nielsen. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 373–380.

[24] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[25] OpenAI. 2023. GPT-4 Model. https://platform.openai.com/docs/guides/gpt

[26] OpenAI. 2023. GPT-4 Technical Report. ArXivabs/2303.08774

[27] Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828* (2023).

[28] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[29] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768* (2023).

[30] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries* 23, 3 (2022), 289–301.

[31] Albrecht Schmidt, Passant Elagroudy, Fiona Draxler, Frauke Kreuter, and Robin Welsch. 2024. Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI. *Interactions* 31, 1 (2024), 24–31.

[32] Qian Wan and Zhicong Lu. 2023. Investigating Semantically-enhanced Exploration of GAN Latent Space via a Digital Mood Board. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–5.

[33] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).

Received 2024