IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 5, October, 2024, pp. 2673-2681
Journal homepage: http://iieta.org/journals/ts

Exploring Multimodal Large Language Models ChatGPT-4 and Bard for Visual Complexity Evaluation of Mobile User Interfaces

Eren Akça[1,2*], Ömer Özgür Tanrıöver[2]

[1] STM Defence Technologies Engineering Inc., Ankara 06530, Turkey
[2] Department of Computer Engineering, Ankara University, Ankara 06830, Turkey

Corresponding Author Email: erenakca88@gmail.com

https://doi.org/10.18280/ts.410540

Received: 29 January 2024
Revised: 29 April 2024
Accepted: 20 August 2024
Available online: 31 October 2024

Keywords:
ChatGPT, Bard, Gemini, LLM, user interface, visual complexity

ABSTRACT

Generative large language models (LLM) are trained for performing natural language processing (NLP) tasks but are known to have emergent properties that can go beyond generating trained text-based language responses. Recently, LLMs have been further augmented with multimodal capabilities such as image annotations and analysis. In this study, we aimed to investigate LLMs in terms of perceptual visual complexity analysis ability through evaluating graphical user interfaces. For this purpose, visual complexity evaluation of user interfaces (UI), which is a non-trivial task, was addressed to explore the possible roles and capabilities of the LLMs in this task. ChatGPT-4 and Bard, two of the most advanced multi modal LLMs, were explored and a comparative evaluation was conducted. According to this exploration, the two LLMs were able to evaluate the visual complexity of different input user interfaces and rank these regarding to their visual complexities. Although LLMs ranking were mostly similar to each other, relatively high differences with the user evaluation-based rankings were observed.

1. INTRODUCTION

Visual complexity has been a difficult concept to define clearly due to its subjective, emotional and perceptual determinants. For this reason, visual complexity perception has for long been investigated and various methods have been developed by the researchers from various disciplines such as psychology [1-3], cognitive science [4-6] and artificial intelligence [7-9]. The concept was first defined by Snodgrass, and Vanderwart [10] as "the amount of details or intricacy in an image". Although it has been tried to be expressed and estimated formally, the proposed approaches are still far from modeling real user perception. For this reason, it has become a matter of curiosity to us whether LLMs; which have produced remarkable human-like responses in many fields, can actually perceive the complexity of UIs similar to users.

ChatGPT and Bard are transformer-based NLP models that can understand and generate general-purpose responses based on large language models, consisting of billions of parameters and trained with a vast amount of dataset. By interacting with humans in their own language (natural language), LLMs may lead to groundbreaking developments in diverse fields ranging from medicine, education, business and finance, law, computer science to education and scientific writing [11]. In 2022, ChatGPT was introduced by OpenAI [12], followed by Meta LLaMA [13] and Google Bard (the name of the LLM was then changed to Gemini) were released in 2023, and later other LLMs such as Claude, Alpaca, etc. have been developed [14]. ChatGPT, Bard, and LLaMA are among the most advanced artificial intelligence tools that pioneer revolutionary developments at the current state of technology. Parallel to the emergence of LLMs, the LLMs has become the focus of many research studies in almost every field. By analyzing the LLM generated responses, ChatGPT and other LLMs were compared based on different criteria [15-17], and their limitations and challenges were investigated [12, 18].

Since conventional LLMs are text-based generative models, studies have commonly analyzed the responses of LLMs to text-based inputs. Studies on how LLMs produce responses to image-based inputs and the accuracy, understandability, consistency and adequacy of these responses are quite limited. Although text-to-image generation studies using large language models [19, 20] are available in the literature, to the best of our knowledge, the effectiveness of LLMs for UI evaluation has not yet been studied thoroughly. With this motivation, our primary research question can be expressed as follows:

RQ1. Can the visual complexity of the UIs be perceived by LLMs?

In addition, the secondary research questions can be listed as below:

RQ2. What factors affect the visual complexity perception of LLMs?

RQ3. Among these factors, are there any other factors that differ from those proposed in previous studies?

RQ4. Does LLM perception of visual complexity level align with user perception level for mobile UIs?

In order to fill the above-mentioned gap and to find answers to these research questions, an attempt is made to investigate the visual complexity perception ability of LLMs based on the

UI images. For this purpose, two advanced LLMs; ChatGPT and Bard, were employed to measure how close the visual complexity perception of the LLMs is to user perception. In addition, LLMs were compared with each other on the issue of UI evaluation, the strengths and areas open to improvement of each LLM were revealed.

The rest of the paper is structured as follows: section II reviews LLMs and their use cases in various fields and the UI visual complexity evaluation methods in sub-sections. In section III, the method to evaluation the role of LLMs in visual complexity analysis for UIs is described. Details of experimental study and test results are given in section IV. In section V, the findings are discussed, followed by the limitations in section VI. Finally, the conclusions are presented in section VII.

2. RELATED WORK

2.1 LLM evaluation studies in different domains

After the release of ChatGPT in 2022, the potential for the use of LLMs has been examined for tasks in numerous research fields. Banerjee et al. [21] tried to measure the reasoning ability of ChatGPT by asking questions about six sub-branches of physiology such as cardiovascular physiology, neurophysiology, endocrine physiology and so on. Accordingly, they statistically analyzed the responses received from ChatGPT to 82 different reasoning questions prepared by the experts. The authors observed that there were significant differences among the responses generated by ChatGPT to the basic concepts and various sub-branches. They also stated that the responses should have a high accuracy rate and be consistent with each other. For this reason, they emphasized that ChatGPT needs further training with more subject-related information such as transfer learning on relevant sub-branches. As an example from software domain, Surameery and Shakor [22], Sobania et al. [23] examined the role of ChatGPT in code debugging and error finding and compared it with software debugging tools. They concluded that ChatGPT may be preferred in terms of cost, speed, and ease of use, but it still needs to be improved in terms of accuracy. Moreover, Biswas [24] examined the potential benefits that ChatGPT can provide to software developers. It was shown that ChatGPT supports the developer in issues such as code completion and correction, predicting and suggesting code snippets, fixing syntax errors, optimizing and refactoring the code and document generation for programming, and can provide a more effective and efficient coding process by saving time.

An early review study [18] examined ChatGPT from wide range of comprehension and reasoning tasks. ChatGPT was found to be successful in producing human-like responses and reasoning in natural language processing tasks. They also mentioned the good performance of ChatGPT's abilities such as anomaly detection, mastery of the subject and communication, in a set of fields such as education, health and industry. However, they also emphasized their reservations on the issues such as ethical, bias and fairness that may cause researchers to be cautious about ChatGPT. They state that these reservations should be addressed, especially in matters such as understanding and interpreting judgments that may vary depending on the person's education, knowledge, background and characteristics.

In addition, Pathak [12] emphasized the prominent features that distinguish ChatGPT from other artificial intelligence products, such as contextual understanding, task compatibility, scalability, ability to produce improved results with the prompting method, and iterative prompting. Plus, they touched on the success of usability, accuracy and reasoning ability in areas such as business and finance, law and legal services, content production, scientific writing, programming and debugging, sales and marketing. They noted that such an advanced product must constantly address personalization, bias and quality control, as well as ongoing challenges such as data privacy and security and adaptation to domain-specific issues. They also pointed out that ChatGPT has limitations such as visual content production, lack of situational awareness, human-level expertise and emotional intelligence.

Hadi et al. [11] argued that LLMs can be used in the fields of health care, financial, engineering and education with the ability to understand the content and produce qualified responses specific to the relevant field without the need for special training. Furthermore, they pointed out that it has some drawbacks such as unintentionally containing biased data, presenting information that does not exist (aka "hallucination"), limited common sense, absence of emotion, and limited domain specific knowledge. In addition, in large language models containing billions of parameters, not knowing exactly what is actually happening at the time of processing and not being able to explain it can be considered among the drawbacks.

In other studies, LLMs were compared with humans based on the responses given to open-ended questions, and it was tried to understand whether LLMs can reason like humans. Duong and Solomon [25] stated that ChatGPT answered questions in the field of genetics as well as humans. However, it could not generate consistent responses to questions requiring critical thinking ability rather than questions requiring memorized knowledge. Moreover, a comparison with benchmark questions from a prestigious competition on computer programming concluded that, contrary to popular belief, humans are significantly better than ChatGPT [26]. On the other hand, Guo et al. [27] tried to understand the characteristics of ChatGPT by analyzing the dataset called HC3 (Human ChatGPT Comparison Corpus), which consists of a large number of questions from various fields such as finance, medicine, legal and physiology, and the responses of ChatGPT and human experts. Accordingly, the important outputs that emerged showed that while ChatGPT provided more detailed and descriptive responses compared to human experts, it could also generate incorrect or misleading information. While there was little sign of emotion in the responses generated by ChatGPT compared to human experts, it was observed that the responses were written in a formal and objective language, without straying from the content. Finally, it was easily understood from the responses that the content was produced by ChatGPT. This was because human experts can convey more result-oriented, short and clear responses with a richer vocabulary and grammar and add emotion and moderate subjectivity.

2.2 LLM comparison studies

Comparative studies of different LLMs were conducted regarding the role, capabilities and limitations of LLMs in various tasks. The responses generated by different LLMs for various tasks were compared with each other. Hadi et al. [11]

compared ChatGPT, Bard and Bing LLMs with each other according to their different features. ChatGPT produced more creative responses, while Bard gave more accurate results. They concluded that Bing is the one that produces the most accurate responses and is user-friendly among these LLMs. In another study, Ahmed et al. [16] evaluated ChatGPT and Bard in terms of accuracy and completeness of the content of the generated responses, integrability into other platforms and performance of human interaction. In this context, although they stated that sufficient standards and measurement metrics have not yet been developed to compare LLMs, they concluded that Bard is better than ChatGPT at interactive dialogues and human-like conversations. However, unlike [11], they stated that ChatGPT produced more accurate results compared to Bard.

In another interesting study, Lozić and Štular [17] measured the potential of LLMs in scientific writing and compared them with each other. According to their results, although ChatGPT, Bing and Bard succeeded in writing relevant articles on the given topics, all LLMs failed in terms of the adequacy of the content of the articles. In particular, this study showed that the reasoning ability of LLMs is still quite limited compared to human researchers. A similar study was conducted by Plevris et al. [28] and LLMs were given various mathematic and logic problems to solve. Accordingly, while it was observed that LLMs could provide fast and satisfactory solutions to basic logic and algebraic questions, they concluded that LLMs were not reliable enough for more complex questions. Another result obtained from the study is that ChatGPT generated more accurate and reasonable results compared to Bard, but the solutions generated by both LLMs are not consistent.

ChatGPT and Bard were also tested and compared with medical information [15, 29]. Many questions regarding various sub-disciplines of medicine were asked to the LLMs, and the consistency, being up-to-date, understandability and accuracy of the responses, as well as reasoning ability of the LLMs, were assessed. In this respect, it has been stated that the responses from LLMs mostly satisfy the human experts on the subject, but LLMs might generate incorrect or illogical responses. They also statistically confirmed that ChatGPT generated more accurate responses than Bard.

2.3 Visual complexity analysis methods

Visual complexity analysis methods for UIs can be examined under two main categories: traditional methods and innovative methods. Traditional methods are based on either direct user evaluation [6, 30-32], combinational metric sets coming from various studies [33-37], or the rules inferred from previous UI evaluation knowledge and experimental studies [38, 39]. In order to evaluate a UI using traditional methods, it is necessary to first understand the factors that affect visual complexity and make visual complexity measurable by transforming these factors into concrete expressions by using metrics or rules. A detailed review of the methods applied to date in the literature has been made by Akça and Tanriöver [40] on the strengths and areas open to improvement of the methods in terms of efficiency regarding to time, cost and performance.

The main advantage of traditional methods is that they produce more consistent and objective results. There are various metrics introduced for this purpose. Some of these metrics are applied for visual complexity analysis through the visible features of UIs such as number of UI elements, element size, alignment, balance, density, grouping, symmetry and so on [33-36]. Conversely, some others are applied through features that are not visible to users but are thought to affect visual complexity such as file size, entropy, compression rate, clusters of colors etc. [37, 41, 42]. The biggest handicap of traditional methods is that it is not known which features of the UI will lead to more accurate and sensitive visual complexity analysis results. Similarly, when analysis is performed using a metric set, it is not known to what extent the metrics contribute to the result. Thus, although traditional visual complexity analysis methods have been the most studied methods in the literature so far, a generalized solution has not yet been revealed.

On the other hand, innovative methods have begun to be implemented over time with the widespread use of machine learning [7, 8] and deep learning [9, 43] techniques, aiming to predict visual complexity without direct human intervention. In these methods, both popular machine learning techniques such as SVM (Support Vector Machine) and popular or specifically developed deep neural networks are used. Unlike traditional methods, in innovative methods, the intermediate steps for visual complexity analysis such as feature extraction, stimulus detection are left to the machine. After training with a sufficiently large dataset, visual complexity analysis is automatically performed by the trained model.

In these approaches, the need to make critical decisions such as which and how metrics and/or rules to be employed for visual complexity analysis is decreased. The better the model learns the more successful, objective and generalizable the visual complexity analysis can be conducted. The key challenge is the necessity of a sufficiently large and diverse dataset and the training cost to create generic deep learning models. Since advanced multimodal LLMs are trained with large and diverse datasets they naturally have potential for visual complexity evaluation task.

Increasing success of deep learning models suggests that use of them to achieve ergonomic, useful, responsive and aesthetic UI designs. LLMs can potentially further help the designer to better understanding the factors affecting the visual complexity of the design under consideration. As reviewed in section 2.1, various studies indicate the potential of LLMs in visual complexity evaluation maybe close to user perception. This is why LLMs are intended to be used in this study. In this sense, the prominent features of LLMs, the reason they are considered are as follows: 1) The capability of producing responses in human language and the potential to clearly state the reasoning behind the responses, 2) The potential to perform visual complexity analysis by considering the factors that might have not been considered in the existing methods. Lastly, to the best of our knowledge, that there are no other studies on visual complexity analysis of UIs using LLMs.

3. METHODOLOGY

In order to answer the research questions that constitute the scope of this study, two different LLMs were first identified according to a set of criteria such as 1) LLM should be multimodal that can understand images, as well as text to process, 2) LLM should not be created for a specific purpose, it should be the one that can be used for general purposes, 3) It should be a baseline LLM that has been studied regarding its capabilities, performance and limits. Besides these criteria, taking into account factors such as the parameter size, task success rates and reasoning ability [17, 43-46], we evaluated that ChatGPT 4.0 and Bard would be suitable for this study among LLMs such as ChatGPT 4.0, Bard, LLaMA, Claude, Bing-Chat and Alpaca.

After deciding on the LLMs, the next decision step was to select and prepare the dataset. The most important factor in determining the dataset was to use one that could represent the widest range of UIs possible. Another concern was that we wanted to use the UI images that had previously been analyzed for visual complexity. Thus, we aimed to use previously obtained results as ground truth for this study. In our previous work, we carried out a study on predicting the visual complexity of mobile UIs using deep learning models [9] trained with mobile UI images selected among the RICO [47] dataset. RICO dataset contains approximately 72000 mobile UI images captured from more than 9000 android applications to be used in scientific research studies such as UI design, UI code generation, user interaction modeling and user perception prediction etc. Inspired by the RICO dataset, another private dataset was created by obtaining UI images from different IOS applications. These images were employed to make comparisons of user evaluations including 98 participants with the results obtained from deep learning models. It was shown that there was a positive correlation between the results. In this study, ten UI images selected from the previously created private dataset were used for LLMs to evaluate visual complexity (see Figure 1).

We then conducted experiments to obtain, evaluate and compare the outputs from LLMs corresponding to the inputs and prompts consecutively. For this purpose, firstly UI images were used as input to two multimodal LLMs. As output, LLMs are expected to produce values by evaluating the visual complexity of the input images and to rank these images according to their visual complexity values. The same procedure was then applied again but this time in addition to the UI images, prompts were also provided as input.

It is known that with corrective prompts it is possible to direct LLMs and obtain more accurate responses. Based on this, the input images were supported with prompting techniques and the LLMs were guided on how to evaluate these images. The responses generated by the LLMs were tried to be improved by iteratively applying prompt engineering methods [11]. While doing this, the LLMs were reminded at regular intervals that the inputs were images of mobile UIs, and they were asked to evaluate the input images by considering like the end user, but it was not stated from what perspective it should evaluate the complexity. The reason for this is to ensure that, acting as an end user, LLM determines how to evaluate the UI images and that LLM discovers which and how the factors affecting visual complexity of the images. Otherwise, LLMs would be subjected to human intervention, so it would be inevitable for them to generate responses that are far from innovative evaluation.

Then, related to the secondary research questions, the success of the LLMs' responses was tried to be measured by using a statistical test. To this end, the rankings formed by LLMs were tested with the Kendall correlation coefficient (aka Kendall's Tau - τ) to get the similarity rate between these rankings. This test was carried out separately for the ranking results created by LLMs and for the results after applying prompts in order to measure the sensitivity of LLMs to prompting. Furthermore, by comparing the factors affecting visual complexity evaluation performed by LLMs with the existing factors defined in previous studies, we aimed to figure out visual complexity factors considered by LLMs.

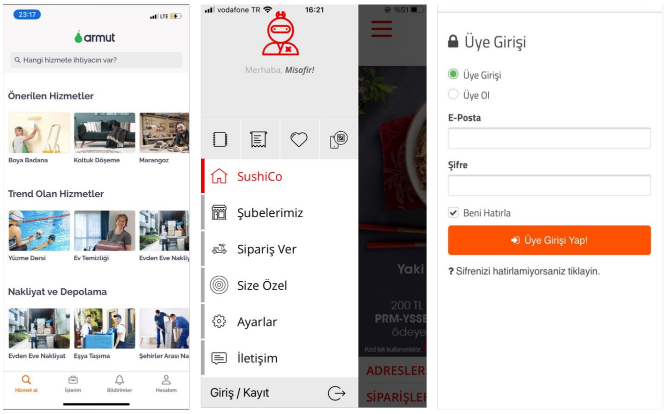


Figure 1. Sample mobile UI images used in the study

The overall systematic of the study was as follows: 1) Ten images were separated into two image sets. 2) Five mobile UI images were given as input to the LLMs. Since Bard cannot be provided with all the images as input at once, the randomly selected images were provided one by one, but this is not the case for ChatGPT-4. 3) After all the input UIs were provided to LLMs, LLMs were asked for ranking the images regarding to its visual complexity evaluation for each image. If the final ranking made by the LLM and the ground truth were not the same, prompting was applied to get the LLM to rank correctly. Accordingly, whether the LLMs can achieve the ranking result, which is considered ground truth, how much prompting is needed, and the success performances of the LLMs are the outputs of this study.

4. EXPERIMENTS

In this section, the details of the study, environmental setup and the results are described. UI images, considered in two separate groups, were evaluated by ChatGPT and Bard. Beforehand, no input was provided to the LLMs that could create a potential bias. During interaction with LLMs, it was ensured that there were no directives or implications in the questions asked. A total of 60 visual complexity evaluation experiments were carried out in 30 independent sessions for each image set. In each session, the output of the LLM was the ranking of the input UI images. Initially, LLMs were explained as follows:

ChatGPT:

"Evaluate and rank the five mobile UI images I will show you based on their visual complexity. Show me the ranking result in a tabular form."

Bard:

"I will show you five mobile UI images step by step. Evaluate the visual complexity for each image and finally rank them."

In our previous study, the mobile UIs in image set 1 were ranked according to the visual complexity evaluation by the human participants as shown in Figure 2. This was used as the baseline to which the output ranking of LLMs were compared. Similarly, the baseline for the image set 2 is shown in Figure 3.

In order to understand the relationship between the rankings made by LLMs and ground truth, a statistical test was used. Since the rankings were compared with each other in our study, there were alternative statistical tests that stood out for this purpose such as Pearson correlation, Spearman rank correlation and Kendall correlation coefficient. We decided to apply the appropriate test by comparing these according to the following reasons. Pearson correlation is mostly suitable for continuous data, however the output of the LLMs were the ranking of the UI images. Hence, the two discrete variables were supposed to be compared. Spearman rank correlation can be applied to ordinal data as well as continuous data. In this respect, it can be said that it is similar to the Kendall correlation coefficient. Nevertheless, the sample size used in this study was small, so the possibility of encountering outliers may be high. To this end, we thought that it was better to apply the test which should be robust enough against outliers. Considering all these arguments, we concluded that the Kendall correlation coefficient may be suitable for this study.

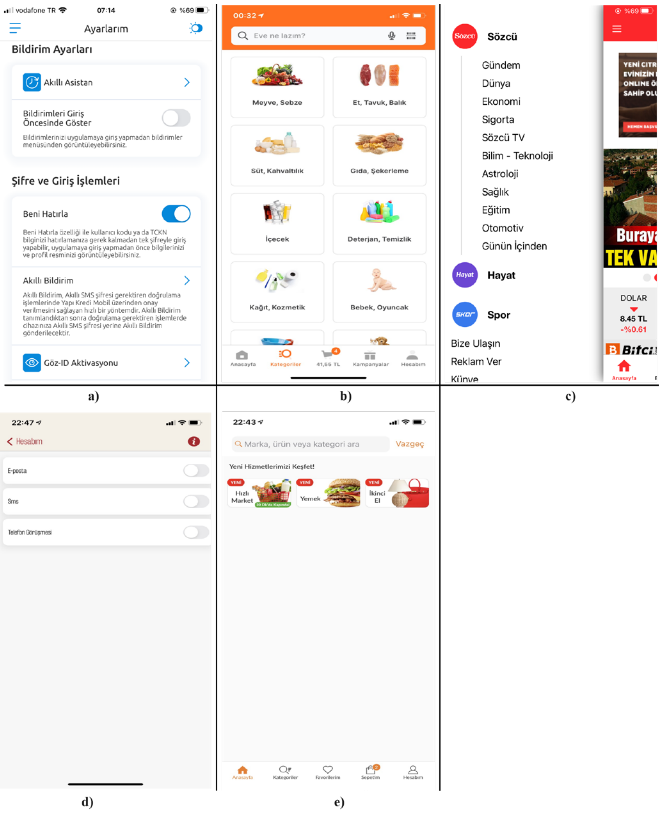


Figure 2. Mobile UI images in image set 1 (the most complex a>b>c>d>e the least complex)

As a result of visual complexity evaluations made by the LLMs, the ranking results are shown in the following tables. The ranking results for ChatGPT and Bard are shown in Table 1 and Table 2, respectively. The results were discussed in two stages: 1) the first ranking results made by the LLMs 2) the modified ranking results as a result of promptings in addition to the first ranking. According to the first ranking results, it can be said that the LLMs produce similar outputs, with the ranking results having a low to moderate level of correlation with the ground truth. However, it is obvious that the variance of ChatGPT's outputs is higher than Bard. That is because unlike Bard's consistent outputs ($0.19 \leq \tau \leq 0.39$), ChatGPT can produce negligible results (τ=0.0), while it can produce results having moderate to strong correlation (τ=0.6) with the ground truth. Moreover, it may also be possible to say that ChatGPT performs visual complexity analysis in a wide perspective ($0.0 \leq \tau \leq 0.6$). When we look at the modified ranking results, it is possible to say that both LLMs showed the expected reaction to the prompts. It has been observed that, with the help of promptings applied, LLMs can produce outputs that have a higher correlation (τ=0.79) with the ground truth. Despite this, another common property of LLMs is that they can remain indifferent to prompts in some case.

Table 1. Ranking results generated by ChatGPT for the mobile UIs in image set 1

Number of Occurrences	First Ranking Results	Ranking Results After Prompting Applied
3	a>c>e>b>d (τ=0.39)	a>b>c>e>d (τ=0.79)
2	b>c>a>e>d (τ=0.39)	No change (τ=0.39)
2	b>c>e>d>a (τ=0.0)	c>b>e>a>d (τ=0.0)
2	a>e>c>b>d (τ=0.19)	a>b>e>c>d (τ=0.6)
2	a>c>b>e>d (τ=0.6)	No change (τ=0.6)

Table 2. Ranking results generated by Bard for the mobile UIs in image set 1

Number of Occurrences	First Ranking Results	Ranking Results After Prompting Applied
3	a>c>d>e>b (τ=0.39)	No change (τ=0.39)
3	b>a>c>e>d (τ=0.31)	a>b>c>e>d (τ=0.79)
2	c>a>d>e>b (τ=0.19)	c>a>b>d>e (τ=0.6)
2	a>c>e>b>d (τ=0.39)	No change (τ=0.39)
2	c>a>b>e>d (τ=0.39)	No change (τ=0.39)

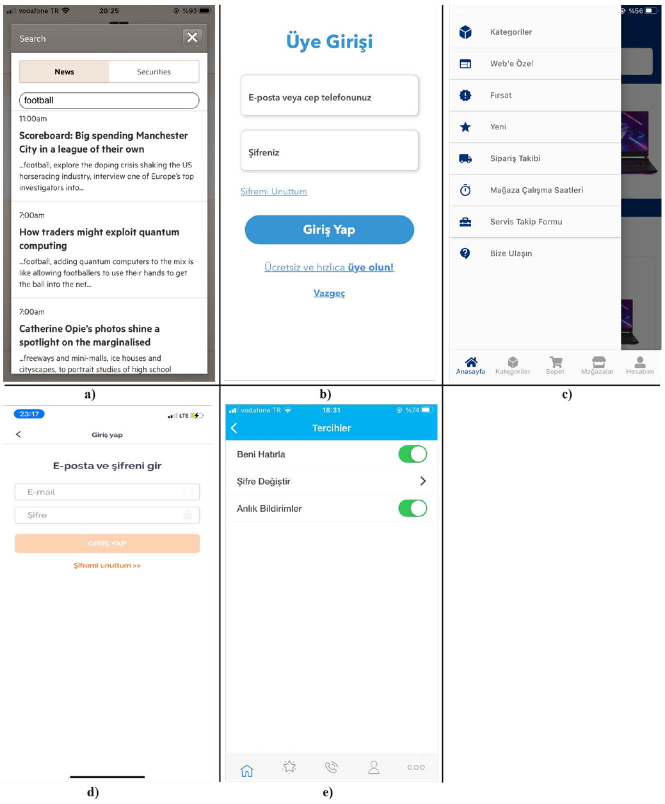
At the end of each session, the LLMs were asked which visual complexity factors were considered. Then, these factors were brought together after all sessions were completed. Accordingly, the factors frequently considered in visual complexity evaluation by the LLMs were as follows:

ChatGPT: Number and order of elements on the UI, text and information density, readability, visual hierarchy, color usage, typography and overall layout.

Bard: Number of elements, element density, color and contrast, typography, whitespace, layout and hierarchy.

The key point here is that LLMs do not take all of these factors into account every time, they evaluate visual complexity with a different set of factors in each session. That is why the ranking order created at the end of each session may differ from each other. However, where similar factors were considered, it was observed that the ranking results were close to each other. The results obtained at the end of each session conducted with LLMs were compared with the ground truth using Kendall correlation coefficient. The resulting correlation coefficient is given in each table together with the ranking results. We also tried visual complexity evaluation with more UIs for each session to obtain a more accurate result. However, we observed that as the number input images increases, the context consistency of the LLMs decreases. That’s why we decided to make comparisons for five UIs.

The mobile UIs in image set 2 are shown in Figure 3. The key difference between two image sets is the UIs in the image set 2 are simpler and text-based compared to the UIs in the image set 1. The ranking results for ChatGPT and Bard for image set 2 are shown in Table 3 and Table 4, respectively.



11:00am
Scoreboard: Big spending Manchester City in a league of their own
...football, explore the doping crisis shaking the US horseracing industry, interview one of Europe's top investigators into...

7:00am
How traders might exploit quantum computing
...football, adding quantum computers to the mix is like allowing footballers to use their hands to get the ball into the net...

7:00am
Catherine Opie's photos shine a spotlight on the marginalised
...freeways and mini-malls, ice houses and cityscapes, to portrait studies of high school

Figure 3. Representation of mobile UI images in image set 2 (the most complex a>b>c>d>e the least complex)

Table 3. Ranking results generated by ChatGPT for the mobile UIs in image set 2

Number of Occurrences	First Ranking Results	Ranking Results After Prompting Applied
3	a>c>d>b>e (τ=0.6)	No change (τ=0.6)
2	a>b>c>d>e (τ=0.99)	No need (τ=0.99)
2	c>a>b>d>e (τ=0.6)	a>c>b>d>e (τ=0.79)
2	c>a>e>b>d (τ=0.19)	c>a>b>e>d (τ=0.39)
2	a>c>d>b>e (τ=0.6)	a>c>b>d>e (τ=0.79)

Table 4. Ranking results generated by Bard for the mobile UIs in image set 2

Number of Occurrences	First Ranking Results	Ranking Results After Prompting Applied
3	c>a>e>b>d (τ=0.19)	a>c>e>b>d (τ=0.39)
2	a>c>d>e>b (τ=0.39)	a>c>b>d>e (τ=0.79)
2	b>e>c>a>d (τ=0.0)	b>e>c>d>a (τ=-0.19)
2	b>e>a>c>d (τ=0.19)	No change (τ=0.19)
2	c>a>e>d>b (τ=0.0)	c>a>b>e>d (τ=0.39)

According to the text-based outputs, it could be said that LLMs are able to understand the aim of the study and produce appropriate responses. Similarly, based on the ranking results and statistical tests when the number of UIs to be compared is kept at a reasonable number, LLMs are able to rank the UI images within the content specified in the question without straying from the topic. However, it has been observed that the ranking results vary considerably. Namely, the resulting tables are sorted by the number of occurrences of the ranking results out of a total of 15 sessions, it has been observed that the rankings are quite different from each other in almost every session. According to the first ranking results for the image set 1, although the LLMs seem to have produced similar results, it can be seen that ChatGPT produced results with a positive correlation (τ=0.6) with ground truth for two times. Additionally, it is possible to say that for the image set 2, this situation becomes more evident in favor of ChatGPT. Because for these images, while ChatGPT frequently produces results that have a strong positive correlation with the ground truth, it is seen that it produces completely the same ranking results as ground truth twice (τ=0.99).

Finally, the responses of the LLMs corresponding to the promptings were also examined. As mentioned before, directions or implications through promptings that could cause potential bias were avoided. Although there are cases where LLMs do not change the rankings they produced at the end of the prompts, it seems that LLMs mostly respond positively to the prompts as clearly seen in all tables. So much so that, with the promptings applied, the correlation rate in the ranking results for both image sets increased positively. Therefore, LLMs are more likely to update their rankings through promptings than insisting on the initial rankings.

5. DISCUSSION

Our study shows that there is a potential to benefit from general purpose multimodal LLMs for visual complexity evaluation of UIs. Although the first ranking results of LLMs has not shown strong correlation with the ground truth, it is possible to obtain results closer to the ground truth when promptings applied. In relation to the RQ2 and RQ3, LLMs do not consider any factors other than the existing visual complexity factors in the literature. Both LLMs perform analysis by considering the basic factors such as the number and order of the UI components, the amount of text and images, color density and overall layout.

There exist variations in the definitions of the concept of visual complexity in the literature. LLMs cannot be expected to consider all the factors affecting it due to their intrinsic reliance on training data source. Therefore, in order for LLMs to evaluate the visual complexity from an alternative perspective, source code of UIs can be given as input in addition to UI images. Considering the rapid development of LLMs, we plan to expand this work by using more UIs with their source codes in a supportive way and LLMs in the future.

Regarding the RQ4, LLMs cannot yet evaluate visual complexity at a level close to user perception unless directional promptings applied. This may be due to the complexity and subjectivity of the problem, but LLMs respond positively with promptings resulting in increased correlation with user perception. In order to improve the performance of LLMs on visual complexity evaluation, a possible future direction might be fine tuning of the LLMs specifically for this problem, so that they can generate more accurate and elaborate responses that can potentially guide designers. Similarly, developing innovative approaches to the prompting techniques can also be considered as potential development direction.

6. LIMITATIONS

There are a number of limitations that may affect the validation of this study. One of these can be thought as the small number of UIs used. It is possible to make more precise measurements using a large number of UI images. Likewise, other LLMs could also be included to extend the evaluation to further compare LLMs with each other to see their capabilities for visual complexity evaluation in a broader perspective. At this point, it could be a good idea to fine-tune a multi modal LLM for this purpose. In this way, more precise observations on the potential drawbacks of existing LLMs can be obtained for UI evaluation.

Additionally, keeping the prompts at a basic level is considered another limitation as more accurate and elaborate responses can be obtained with advanced prompting techniques. However, we plan to examine the further responses when we knowingly and willingly direct LLMs and make use of various prompting techniques. Another value-adding approach could be trying to understand the internal mechanism and the decision arguments of the LLMs with explainable AI techniques. Since this type of research requires a large amount of effort, we plan to a study this issue in the future. In this way, we may better understand the reasoning and maybe direct LLMs to take initiative and to discover new factors that affect visual complexity.

7. CONCLUSION

The purpose of this study was to reveal the potential of LLMs for visual complexity understanding. For this purpose, the role and potential of LLMs for UI visual complexity analysis were examined through an experimental study. With the evaluation study conducted with ChatGPT and Bard, it was observed that although both LLMs were able to generate meaningful responses, there were still discrepancies between the responses obtained by user evaluations. Although the responses of LLMs can be improved with the directional promptings, it is obvious that LLMs are still open to improvement in terms of visual complexity understanding.

REFERENCES

[1] Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), A Source Book of Gestalt Psychology. Kegan Paul, Trench, Trubner & Company, pp. 71-88. https://doi.org/10.1037/11496-005

[2] Leeuwenberg, E.L. (1971). A perceptual coding language for visual and auditory patterns. The American Journal of Psychology, 84: 307-349. https://doi.org/10.2307/1420464

[3] Patel, L.N., Holt, P.O. (2000). Modelling visual complexity using geometric primitives. Orlando: Proceedings, Systemics, Cybernetics and Informatics.

[4] Mack, M.L., Oliva, A. (2004). The perceptual dimensions of visual simplicity. Journal of Vision, 4(8): 719. https://doi.org/10.1167/4.8.719

[5] Olivia, A., Mack, M.L., Shrestha, M., Peeper, A. (2004). Identifying the perceptual dimensions of visual complexity of scenes. In Proceedings of the Annual Meeting of the Cognitive Science Society, Chicago, USA, 26(26): 1041-1046.

[6] Harper, S., Michailidou, E., Stevens, R. (2009). Toward a definition of visual complexity as an implicit measure of cognitive load. ACM Transactions on Applied Perception (TAP), 6(2): 1-18. https://doi.org/10.1145/1498700.1498704

[7] Wu, O., Hu, W., Shi, L. (2013). Measuring the visual complexities of web pages. ACM Transactions on the Web (TWEB), 7(1): 1-34. https://doi.org/10.1145/2435215.2435216

[8] Mao, Y. (2019). User interface evaluation with machine learning methods. Dissertation, University of Michigan.

[9] Akça, E., Tanrıöver, Ö.Ö. (2022). A deep transfer learning based visual complexity evaluation approach to mobile user interfaces. Traitement du Signal, 39(5): 1545-1556. https://doi.org/10.18280/ts.390511

[10] Snodgrass, J.G., Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6(2): 174-215. https://psycnet.apa.org/doi/10.1037/0278-7393.6.2.174

[11] Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Hassan, S.Z., Shoman, M., Wu, J., Mirjalili, S., Shah, M. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprints.

[12] Pathak, A. (2023). Exploring chatgpt: An extensive examination of its background, applications, key challenges, bias, ethics, limitations, and future prospects. Applications, Key Challenges, Bias, Ethics, Limitations, and Future Prospects. https://dx.doi.org/10.2139/ssrn.4499278

[13] Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.A., Lacroix T., Rozière B., Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G., (2023). Llama: Open and Efficient Foundation Language Models. arXiv Preprint arXiv: 2302.13971. https://doi.org/10.48550/arXiv.2302.13971

[14] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R. (2023). A survey of large language models. arXiv Preprint arXiv: 2303.18223. https://doi.org/10.48550/arXiv.2303.18223

[15] Yıldız, M.S. (2023). Comparing response performances of Chatgpt-3.5, Chatgpt-4 and bard to health-Related questions: Comprehensiveness, accuracy and being up-to-date. Chatgpt-4 and Bard to Health-Related Questions: Comprehensiveness, Accuracy and Being Up-to-Date. https://doi.org/10.2139/ssrn.4503443

[16] Ahmed, I., Roy, A., Kajol, M., Hasan, U., Datta, P.P., Reza, M.R. (2023). ChatGPT vs. Bard: A comparative study. Authorea Preprints. UMBC Student Collection. https://doi.org/10.22541/au.168923529.98827844/v1

[17] Lozić, E., Štular, B., (2023). ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI LLMs at scientific writing? Preprint 2023, v.3. https://doi.org/10.48550/arXiv.2309.08636

[18] Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A., Latif, S. (2023). Exploring ChatGPT capabilities and limitations: A critical review of the NLP game changer. Preprints 2023: 2023030438. https://doi.org/10.20944/preprints202303.0438.v1

[19] Liu, V., Chilton, L.B. (2022). Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pp. 1-23. https://doi.org/10.1145/3491102.3501825

[20] Oppenlaender, J. (2023). A taxonomy of prompt modifiers for text-to-image generation. Behaviour & Information Technology, 1-14. https://doi.org/10.1080/0144929X.2023.2286532

[21] Banerjee, A., Ahmad, A., Bhalla, P., Goyal, K. (2023). Assessing the efficacy of ChatGPT in solving questions based on the core concepts in physiology. Cureus, 15(8): e43314. https://doi.org/10.7759/cureus.43314

[22] Surameery, N.M.S., Shakor, M.Y. (2023). Use chat GPT to solve programming bugs. International Journal of Information Technology and Computer Engineering, 31: 17-22. https://doi.org/10.55529/ijitc.31.17.22

[23] Sobania, D., Briesch, M., Hanna, C., Petke, J. (2023). An analysis of the automatic bug fixing performance of chatgpt. In 2023 IEEE/ACM International Workshop on Automated Program Repair (APR), Melbourne, Australia, pp. 23-30. https://doi.org/10.1109/APR59189.2023.00012

[24] Biswas, S. (2023). Role of ChatGPT in computer programming. Mesopotamian Journal of Computer Science, 2023: 9-15. https://doi.org/10.58496/MJCSC/2023/002

[25] Duong, D., Solomon, B.D. (2024). Analysis of large-language model versus human performance for genetics questions. European Journal of Human Genetics, 32(4): 466-468. https://doi.org/10.1038/s41431-023-01396-8

[26] Koubaa, A., Qureshi, B., Ammar, A., Khan, Z., Boulila, W., Ghouti, L. (2023). Humans are still better than ChatGPT: Case of the IEEE Xtreme competition. Heliyon, 9(11): e21624. https://doi.org/10.1016/j.heliyon.2023.e21624

[27] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv Preprint arXiv: 2301.07597. https://doi.org/10.48550/arXiv.2301.07597

[28] Plevris, V., Papazafeiropoulos, G., Rios, A.J. (2023). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. arXiv Preprint arXiv: 2305.18618. https://doi.org/10.48550/arXiv.2305.18618

[29] Patil, N.S., Huang, R.S., van der Pol, C.B., Larocque, N. (2024). Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Canadian Association of Radiologists Journal, 75(2): 344-350. https://doi.org/10.1177/08465371231193716

[30] Geissler, G.L., Zinkhan, G.M., Watson, R.T. (2006). The influence of home page complexity on consumer attention, attitudes, and purchase intent. Journal of Advertising, 35(2): 69-80. https://doi.org/10.1080/00913367.2006.10639232

[31] Tuch, A.N. (2007). Visual complexity of websites and its effects on experiential, psychophysiological, visual search reaction time and recognition responses. Doctoral Dissertation, Department of Cognitive Psychology and Methodology, University of Basel.

[32] Michailidou, E., Harper, S., Bechhofer, S. (2008). Visual complexity and aesthetic perception of web pages. In Proceedings of the 26th Annual ACM International Conference on Design of Communication, pp. 215-224. https://doi.org/10.1145/1456536.1456581

[33] Purchase, H.C., Hamer, J., Jamieson, A., Ryan, O. (2011). Investigating objective measures of web page aesthetics and usability. In Proceedings of the Twelfth Australasian User Interface Conference, Perth Australia, 117: 19-28.

[34] Taba, S.E.S., Keivanloo, I., Zou, Y., Ng, J., Ng, T. (2014). An exploratory study on the relation between user interface complexity and the perceived quality. In Web Engineering: 14th International Conference, ICWE 2014, Toulouse, France. Proceedings. Springer International Publishing. Springer, Cham, 14: 370-379. https://doi.org/10.1007/978-3-319-08245-5_22

[35] Alemerien, K.A. (2014). Metrics and tools to guide design of graphical user interfaces. Doctoral Dissertation, North Dakota State University.

[36] Riegler, A., Holzmann, C. (2018). Measuring visual user interface complexity of mobile applications with metrics. Interacting with Computers, 30(3): 207-223. https://doi.org/10.1093/iwc/iwy008

[37] Oulasvirta, A., De Pascale, S., Koch, J., Langerak, T., Jokinen, J., Todi, K., Laine, M., Kristhombuge, M., Zhu, Y., Miniukovich, A., Palmas, G., Weinkauf, T., (2018). Aalto interface metrics (AIM): A service and codebase for computational GUI evaluation. In: 31st Annual ACM Symposium on User Interface Software and Technology. Adjunct Proceedings, New York, USA, pp. 16-19. https://doi.org/10.1145/3266037.3266087

[38] Ines, G., Makram, S., Mabrouka, C., Mourad, A. (2017). Evaluation of mobile interfaces as an optimization problem. Procedia Computer Science, 112: 235-248. https://doi.org/10.1016/j.procs.2017.08.234

[39] Soui, M., Chouchane, M., Mkaouer, M.W., Kessentini, M., Ghedira, K. (2020). Assessing the quality of mobile graphical user interfaces using multi-objective optimization. Soft Computing, 24: 7685-7714. https://doi.org/10.1007/s00500-019-04391-8

[40] Akça, E., Tanriöver, Ö.Ö. (2021). A comprehensive appraisal of perceptual visual complexity analysis methods in GUI design. Displays, 69: 102031. https://doi.org/10.1016/j.displa.2021.102031

[41] Bakaev, M., Heil, S., Khvorostov, V., Gaedke, M. (2018). HCI vision for automated analysis and mining of web user interfaces. In Web Engineering: 18th International Conference, ICWE 2018, Cáceres, Spain, Proceedings Springer International Publishing. Springer, Cham, 18: 136-144. https://doi.org/10.1007/978-3-319-91662-0_10

[42] Boychuk, E., Bakaev, M. (2019). Entropy and compression based analysis of web user interfaces. In Web Engineering: 19th International Conference, ICWE 2019, Daejeon, South Korea, Proceedings. Springer International Publishing. Springer, Cham, 19: 253-261. https://doi.org/10.1007/978-3-030-19274-7_19

[43] Dhengre, S., Mathur, J., Oghazian, F., Tan, X., McComb, C. (2020). Towards enhanced creativity in interface design through automated usability evaluation. In 11th International Conference on Computational Creativity, Coimbra, Portugal, pp. 366-369.

[44] Wu, S., Koo, M., Blum, L., Black, A., Kao, L., Scalzo,

F., Kurtz, I. (2023). A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. arXiv Preprint arXiv: 2308.04709. https://doi.org/10.48550/arXiv.2308.04709

[45] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A. (2023). A comprehensive overview of large language models. arXiv Preprint arXiv: 2307.06435. https://doi.org/10.48550/arXiv.2307.06435

[46] Sandmann, S., Riepenhausen, S., Plagwitz, L., Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nature Communications, 15(1): 2050. https://doi.org/10.1038/s41467-024-46411-8

[47] Deka, B., Huang, Z., Franzen, C., Hibschman, J., Afergan, D., Li, Y., Nichols, J., Kumar, R. (2017). Rico: A mobile app dataset for building data-driven design applications. In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. Quebec City, Canada, pp. 845-854. https://doi.org/10.1145/3126594.3126651