Using ChatGPT for Human–Computer Interaction Research: A Primer

March 2023

Wilbert Tabone (ORCID: 0000-0002-5796-9571) & **Joost de Winter** (ORCID: 0000-0002-1281-8200)

Department of Cognitive Robotics, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, The Netherlands

Abstract

Qualitative data analysis research is often a time-consuming process that involves coding and categorizing large amounts of text, such as questionnaire responses and interview responses. Moreover, the analysis of text data has faced criticism for a lack of replicability. We examined whether ChatGPT can be used as a valid tool in human–computer interaction research by applying it to (1) textbox questionnaire responses about augmented-reality interfaces for pedestrian-vehicle interaction, (2) interview data obtained from a study in which participants in the role of pedestrians experienced interfaces in an immersive virtual simulator, (3) transcribed think-aloud data of participants who viewed a real painting and its replica. A hierarchical approach was used, in which the ChatGPT API was tasked with producing scores or summaries of text batches, which were subsequently aggregated. Results showed that (1) ChatGPT generated sentiment scores that correlated highly with results previously obtained from numeric rating scales. Additionally, (2) by inputting automatically-transcribed interviews into ChatGPT, it provided meaningful summaries of the strengths and weaknesses of the interfaces. Furthermore, (3) ChatGPT's summary of the think-aloud data successfully highlighted subtle differences in the participants' statements between the real painting and the replica, which would have been difficult to identify manually. In conclusion, ChatGPT is a viable tool for analyzing text data in HCI research.

Keywords: Prompt engineering, human-subject research, Application Programming Interface (API), Replicability.

1. Introduction

OpenAI's ChatGPT has taken the research world by storm, garnering media attention and sparking debates. Released in November 2022, ChatGPT has shown great promise in various areas, such as interpreting computer code (Sobania et al., 2023; Tate et al., 2023), generating prompts for input into art-generating tools (Pavlik, 2023), document writing and translation (Tate et al., 2023), text interpretation and review (Zhang & Simeone, 2022; Zhong et al., 2023), generating creatively written text such as poetry (Kirmani, 2023), writing hospital discharge summaries (Patel & Lam, 2023), and as an assistive tool in education (Bommarito & Katz, 2022; Gao et al., 2022), amongst many other applications. Although ChatGPT offers many opportunities, others have highlighted several risks, including the potential for plagiarism in academic writing (González-Padilla, 2022; Krukar & Dalton, 2020; Thorp, 2023) and education (De Winter, 2022; Gilson et al., 2022; Rudolph et al., 2023; Stoker-Walker, 2022), generation of incorrect computer code (Vincent, 2022) or incorrect mathematical solutions (Frieder et al., 2023), and the possibility of inaccurate or biased output (Alba, 2022; Borji, 2023; Van Dis et al., 2023).

In this work, we explored whether ChatGPT can be applied as a valid tool in Human–Computer Interaction (HCI) research. HCI is the field that focuses on the design and use of computer technology and examines interfaces between people and computers. HCI research may involve various types of methods, including questionnaires, interviews, psychophysics methods, virtual reality setups, and field tests. In this process, the HCI researcher often collects text data. For example, in questionnaire research, a number of free-response items may be included to allow the respondents to reflect on a specific type of HCI design, while in interviews, verbal data is collected that is later transcribed and analyzed (Gerlach & Kuo, 1991; Gubrium & Holstein, 2001; Maraj et al., 2016; Schelble et al., 2022), and in field studies or VR studies, a think-aloud protocol is often used to enrich the data collection (Clemmensen & Roese, 2010; Kjeldskov & Skov, 2003; Zhang & Simeone, 2022; Zhao & McDonald, 2010). The analysis of text-based data has been criticized for lacking replicability (Chakrabarti & Frye, 2017; Kitto et al., 2023; Krippendorff, 2004). For example, although the widely employed thematic analysis approach (Braun & Clarke, 2006; De Carvalho & Fabiano, 2021; Kiger & Varpio, 2020) has been praised for being able to reveal underlying themes in texts, whether other researchers can replicate these themes has been an ongoing source of debate (Roberts et al., 2019). ChatGPT has the potential to offer new possibilities for processing and interpreting text-based data.

Here, we investigated whether ChatGPT could be used as a reliable tool in HCI research by analyzing questionnaire responses, interviews, and think-aloud data that were collected in three prior studies (De Winter et al., 2022; Tabone et al., 2023a, 2023b). Our objective was to determine whether the results provided by ChatGPT would be comparable, or superior, to the results generated by the originally conducted manual analyses and whether ChatGPT could be employed as an assistant tool in HCI research. Three separate experiments were performed using either the ChatGPT API (a feature released on March 1, 2023) or the online website, where it was applied to questionnaire data, interview data, and transcribed think-aloud data.

2. Study 1: Questionnaire Textbox Data

In Tabone et al. (2023a), 992 respondents rated nine new augmented reality (AR) interfaces for pedestrian-vehicle interaction. Each interface was presented in two videos depicting a crossing situation in a virtual environment, where a single vehicle approached from the right. In each video, the interface was presented in a yielding or non-yielding state, with the former communicating that the approaching vehicle would stop for the pedestrian and the latter communicating the opposite. Figure 1 provides still frames from the videos as an example. As part of the questionnaire, respondents had to complete several rating scales related to the interface's intuitiveness and convincingness in communicating the message to cross or not to cross the road. Ratings regarding acceptance, attractiveness, aesthetics, ease of understanding, and the adequacy of information were also presented. A free-text area was added at the end of the questionnaire, allowing respondents to elaborate on their ratings: "*Please add a few words to justify your choices above (e.g., comment on the shape, colour, functionality, and the clarity of the interface)*".

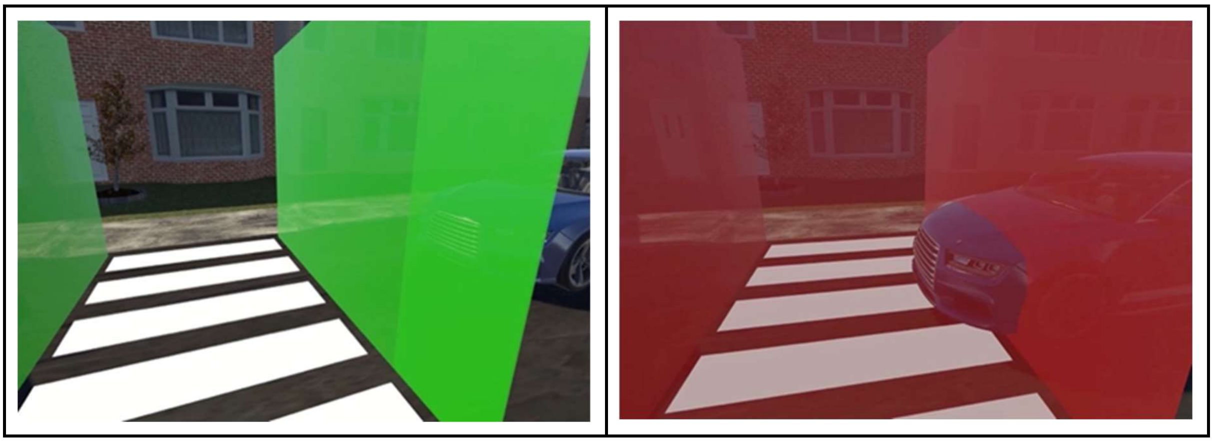

Figure 1. Example of frames from the videos presented in the online questionnaire. In this instance, the ‘Virtual fence’ (interface number 6) interface is seen in its yielding (left) and non-yielding states (right) (Tabone et al., 2023a).

In the original study, the quantitative ratings for intuitiveness, convincingness etc. were statistically analyzed and combined into a composite score, the meaning of which can be characterized as ‘whether the AR interface is good or not’ (Tabone et al., 2023a). Additionally, the text-based responses were thematically analyzed using a manual procedure of reviewing all responses and selecting ‘interesting’ responses for a separate document. From an average of 46 selected text comments per interface, a subset of 4 comments (2 with a positive sentiment, and 2 with a negative sentiment) that were deemed to represent the major themes of that particular concept were selected and presented in the publication.

The qualitative data analysis was a laborious task that took a lot of time, and a limitation was that the researcher’s subjectivity may have influenced the analysis. Moreover, the manual method used was insufficient to make various comparisons. More specifically, it was difficult to assess whether the results from the thematic analysis were somehow linked to the calculated composite score.

2.1. Methods

In the new analysis, all text responses were extracted from the raw questionnaire data, available from the online supplementary material of Tabone et al. (2023a). This yielded 9 columns (1 for each interface) and 992 rows, for a total of 8,928 entries. The data was unaltered, and the responses were submitted to the ChatGPT API (date: 4 March 2023, model: gpt-3.5-turbo). The parameter governing randomness, referred to as ‘temperature’, was set to a value of 0. This configuration may be advantageous in certain tasks, such as sentiment analysis, where the absence of random variation is preferred. This attribute also highlights the benefits of using the ChatGPT API, in contrast to the earlier web interface, which does feature randomness.

Of note, the raw data submitted to ChatGPT contained not only meaningful statements, but also plenty of meaningless statements, staccato text, or text in different languages (see Table 1 for a random selection of 20 comments for one of the nine interfaces). Thus, we tested whether ChatGPT could validly handle raw text input.

Table 1. Selection of comments about the ‘Virtual fence’ interface (20 of 992 comments were selected using a random number generator)

Interesting and very good.
the would be fun to test in the future
I have no idea

Really clear, but a bit too much
Full colour very clear
easy to see with the bright colours
Best one so far...
Not so great grapic quality
To big sign
i have no comment
On det kommer någon gående då stannar man eftersom det är det det betyder.
extremely clear way to show when it is safe and unsafe to cross
big, bold and very clear
Bruce Springsteen is a communist
Want to see by myself if it’s safe
felt more intuitive than the second option but i still prefer the first option
The interface is too big
It is hard to trust.
I think the fencing is a good way to signal whether or not it's safe to cross the street. However, the non-yielding fence isn't as effective at getting the message across, apart from the colour, as the pedestrian crossing is still open. I think it could be confusing, especially to someone who's colourblind. Maybe a big X to cross out the pedestrian crossing/obscure the path would be more effective here.
niet van toepassing
It was a bit confusing to show a crossing if you were not supposed to cross. Even though it was red I still saw the crossing and thought it was maybe ok.

Since there are limits to the amount of text that could be submitted at any one time, it was decided to adopt a ‘hierarchical approach’ to data processing, where batches of text were submitted, and the numeric outputs were subsequently averaged. Specifically, each column was automatically split into nine sections of ~100 rows each. Each section was then separately submitted to the API, with a specific prompt:

“*The comments below were obtained from respondents in an online survey about an AR interface in a road-crossing scenario. If you read the comments below, what is the quality of the interface on a scale from 1 = bad to 10 = good? Only report a number between 1 and 10. Round to two decimals*”, followed by the 100 comments, each on an individual line.

After all the batches were submitted, the mean sentiment score for each interface was calculated by averaging the sentiment scores of its 10 batches. The correlation coefficient was computed between the mean sentiment score of the interfaces and the composite score of the interfaces previously reported by Tabone et al. (2023a).

Next, we examined whether the prompt could be improved by trial and error. More specifically, we repeated the above process using different prompts, in an attempt to maximize the correlation coefficient between the mean sentiment score obtained through ChatGPT, and the mean composite score (Tabone et al., 2023a).

2.2. Results

The ChatGPT API provided outputs of the 90 batches (i.e., 9 interfaces x 10 batches per interface) in a total of 72 seconds on a standard laptop and internet connection, which compares favorably to the time taken by the human authors to go over all 8928 comments manually, taking an estimated 8 to 10 hours of work.

The first prompt we tried (see Methods) yielded a very strong correlation between the ChatGPT outcomes and the composite from the numeric rating scales (r = 0.975). In other words, ChatGPT produced sentiment scores that correlated extremely highly with aggregated human ratings.

Using trial and error we altered the prompt into the following: "*Looking at the comments, score the interface, from 1 to 100. Only report a number between 1 and 100, rounded to two decimals*". The correlation coefficient between the two measures was found to be close to 1 (r = 0.99). The scatter plot in Figure 2 shows the mean sentiment score from the ChatGPT outputs against the composite score from the online questionnaire study.

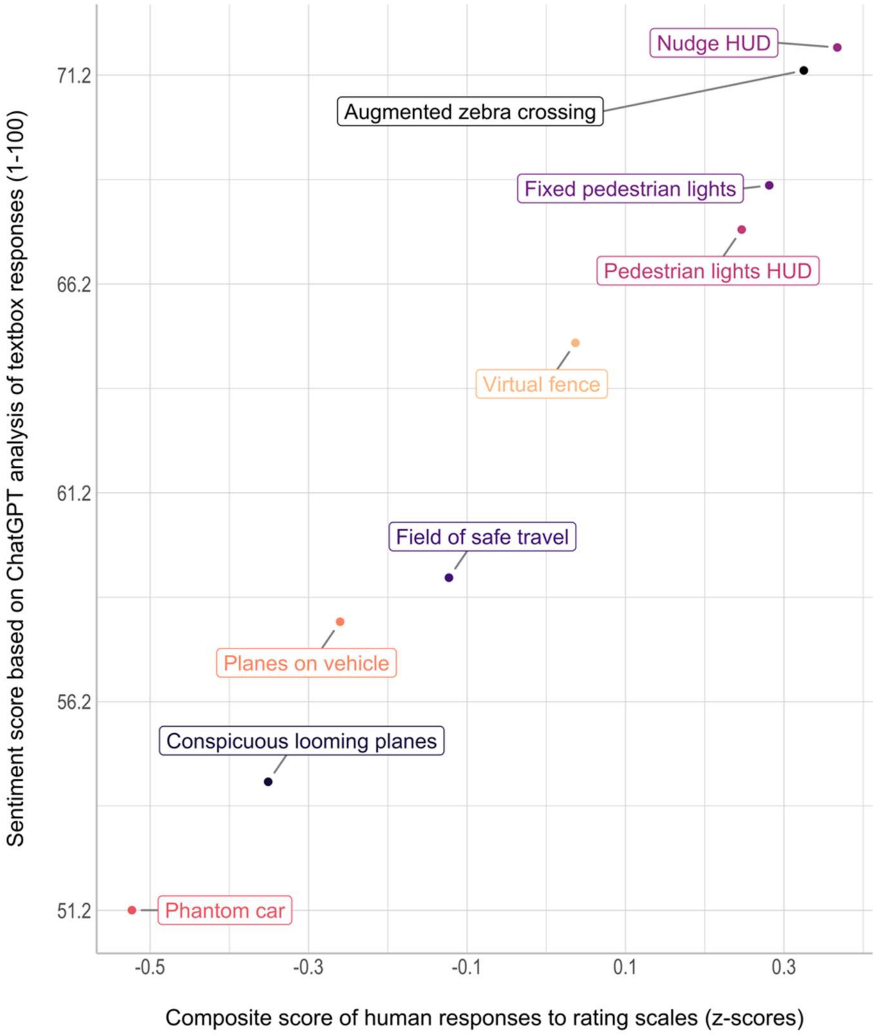

Figure 2. Scatter plot showing the mean sentiment score calculated from the ChatGPT outputs against the composite score based on rating scales in the online questionnaire study (extracted from Tabone et al., 2023a). The composite score has a mean of 0 and a standard deviation of 1 in the population of respondents.

It is noteworthy that the final prompt is very short, and devoid of context. One thing we discovered in the trial-and-error process was that longer prompts, which provided more context (such as that the experiment involved AR and a car and a pedestrian in a road crossing scenario) were not needed for achieving a high correlation. It was also not needed to mention the dimension across which the sentiment score should be provided (e.g., quality, clarity).

Furthermore, we explored the impact of batch size and randomness parameter on sentiment analysis results. We found that varying the batch size (e.g., using batches of 25 instead of 100 comments) or adjusting the randomness parameter (by setting it to a higher value than 0 and averaging the results of multiple repetitions) did not significantly affect the correlation coefficient.

Figure 3 illustrates this finding. We generated 100 sentiment scores for the first batch of comments using different levels of 'temperature' (the setting can range from 0 to 2 in the OpenAI API). We then calculated the mean and standard deviation across these sentiment scores. As shown in Figure 3, the mean remained relatively constant across different temperature settings, while the standard deviation increased with higher temperature values. This indicates that the sentiment analysis results remained unbiased regardless of the temperature value used. Therefore, for the present sentiment analysis task, there is no specific value that requires a nonzero temperature setting.

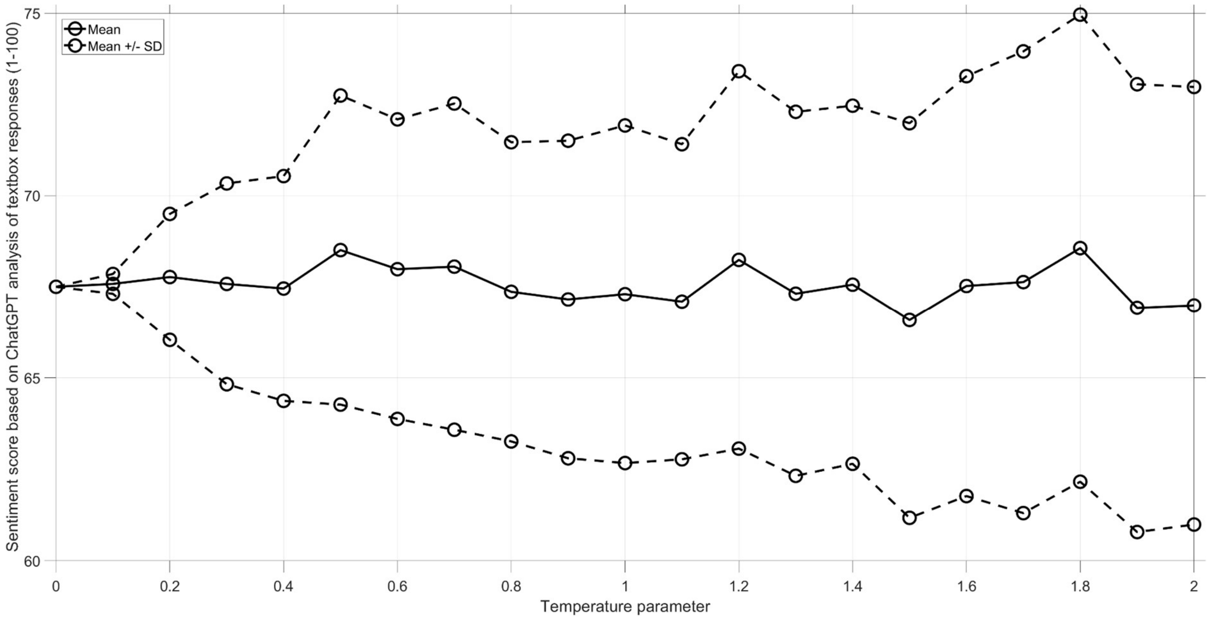

Figure 3. Mean and mean ± standard deviation of the sentiment score based on ChatGPT analysis of textbox responses, for one batch of 100 responses, as a function of the temperature parameter (in increments of 0.1).

3. Study 2: Interview Data

In Tabone et al. (2023b), the effect of nine AR interfaces, as well as a control condition without AR interface, on pedestrian crossing behavior was examined in a virtual simulator environment. The experiment was a follow-up of the online study described in the previous section and was conducted at the University of Leeds using the Highly Immersive Kinematic Experimental Research (HIKER) simulator (Figure 4). The experiment was approved by the University of Leeds Research Ethics Committee under ethics reference number LLTRAN-150. All participants provided written informed consent.

The experiment consisted of 120 trials per participant, each being a combination of one of the interfaces, the location of the visual distraction, and yielding or non-yielding vehicle. The 120 trials were presented in a counterbalanced block design with 10 blocks (one block per interface). After each block, the participant was interviewed for about 3 minutes, starting by asking whether they were comfortable, and then asking "*what did you think of this particular interface/situation?*", and depending on the participant's response, followed by questions about the clarity of the meaning of the color coding (e.g., "*is the color code making any sense to you?*") and a comparison between states (e.g., "*between the two states, were there any preferences*").

Processing the interview data would be a challenging task, which may involve lots of transcription work as well as subjective interpretation. Therefore, it was decided to transcribe the audio-recorded interviews automatically and submit the transcripts to ChatGPT to examine whether ChatGPT could automatically provide insights into the strengths and weaknesses of each interface.

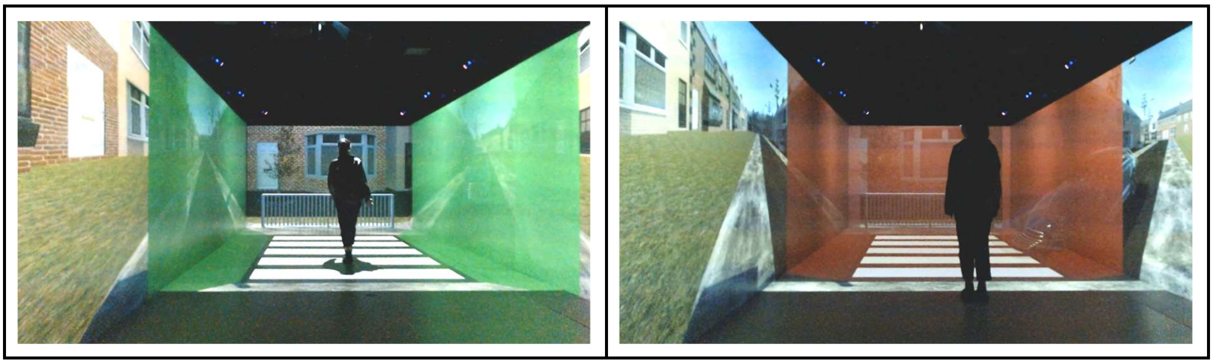

Figure 4. Yielding (left) and non-yielding (right) state of the *Virtual fence* interface.

3.1. Methods

First, the 300 voice recordings were transcribed by submitting the audio files to the paid online service Otter.ai (2023). An example of a transcript for the *Virtual Fence* is shown in Figure 5. As can be seen, the transcripts do not distinguish between the interviewer and interviewee, and contain some transcription errors.

Next, each anonymous transcript was submitted to the ChatGPT API (date: 4 March 2023, model: gpt-3.5-turbo) through a custom script. The following prompt was used: "*Looking at the participant's responses, provide a summary of the interface*", followed by the participant's interview on the subsequent lines.

After collecting all responses from all 300 interviews, the responses per interface were again submitted to ChatGPT. The same prompt was used: "*Looking at the participant's responses, provide a summary of the interface*", including all 30 summaries from the participants.

00:00
So, say 2.8 and of block five. Do you feel comfortable?

00:06
Still feel okay. Yeah.

00:08
Okay, so how to evaluate this interface?

00:12
I'm certainly uncomfortable with the designing. I mean, because of you have like a blocking both sides and I can't really see clearly what's happening behind the color color. And then with the crossing lanes is kind of disturbing. Yeah, in terms of the designing is kind of disturbing to the eye

because I prefer to look natural than having like too much digitalization on ice. Where the

00:43
starting point, influence your decision. Like, when you look into different directions. The interface with this turbine all the time, oh, it's somehow less disturbing when you are facing a specific direction.

01:01
It doesn't matter. It doesn't matter which direction I'm facing. But it I feel disturbing with the designs. I'm not preferring to have like, full color coded around my eyes.

01:14
Okay, so is the color code making any sense to you? Very clearly indicates its meaning.

01:25
I'd say the color choosing is indicating correct meaning, but I don't prefer like to have like a full block. Designing throughout my viewing. Yeah.

01:39
Okay, so do you comfortable with the choice, the confidence in the choice you're making with this interface?

01:48
Because if I'm slightly uncomfortable with designing, it doesn't mean that I didn't trust the decision by interfaces. I'd say, Oh, I see. I still have have a trust. It's just that I don't feel really comfortable. Okay, that's good. Thank you.

Figure 5. Example of an automated transcription (Participant 8, Virtual Fence), created by Otter.ai (2023).

3.2. Results

The result was a summary of the qualities of each interface (see Tabone et al., 2023a, for all summaries). Our script, which connected to the ChatGPT API, took a total of 25 min to process all 300 interviews, including creating the overall summaries.

The overall summary for the *Virtual fence* is provided below. The summary provides useful insights that designers may use to improve the interface. For example, it mentions occlusion effects, that is, although participants felt safe inside the green tunnel, a problem is that the pedestrian may not be able to see the car anymore. In conclusion, we have shown that interview data can be transcribed and summarized automatically.

Overall, the participants had mixed feelings about the interface. Some found it clear and easy to use, while others found it confusing and overwhelming. The green and red states were generally well-received, with participants feeling safer and more confident when the green state was active. However, some participants had concerns about the interface obscuring their view of the road and potentially causing accidents. The zebra crossing was generally appreciated, but some participants found it confusing and potentially misleading. Accessibility features were also noted as a positive aspect of the interface. Overall, there were varying opinions on the effectiveness and usability of the interface.

4. Experiment 3: Think Aloud

In this study, data from a think-aloud protocol was recorded while participants, wearing a mobile eye tracker, freely gazed at Rembrandt's *The Night Watch* (1642) in Amsterdam's Rijksmuseum (De Winter et al., 2022). The aim of the study was primarily to assess the participants' attentional distribution for the work of art. The analysis of the gaze data was supported by the recorded statements from the participants as they freely spoke out their thoughts while gazing at the painting.

The recorded voice data (in Dutch) was previously transcribed and analyzed using a hybrid approach, by combining thematic analysis (Braun & Clarke, 2006) with a quantitative method of tabulating word frequencies and connectedness (Heikoop et al., 2018). The main themes were identified, and the number of times was counted a target word belonging to each theme was spoken.

A follow-up experiment was conducted to assess the robustness of the first experiment, using a replica of the painting in a laboratory setting and with new participants. The participants in the Rijksmuseum study were recruited by contacting acquaintances of the authors and television crew and by inviting (ex-)students, whereas participants in the replica study were recruited from the student population and the teaching and administration staff of the Delft University of Technology. The task instructions and experiment duration (5 min of viewing per participant) were identical between the two settings (Figure 6).

The authors discovered that admiration-related words (i.e., beautiful, cool, enjoy, fantastic, fascinating, great, impressive, nice, special, splendid, unique, wonderful, wow) were uttered more frequently for the real painting compared to the replica painting, other comparisons of think-aloud data between the two versions of the painting did not yield statistically significant differences (De Winter et al., 2022).

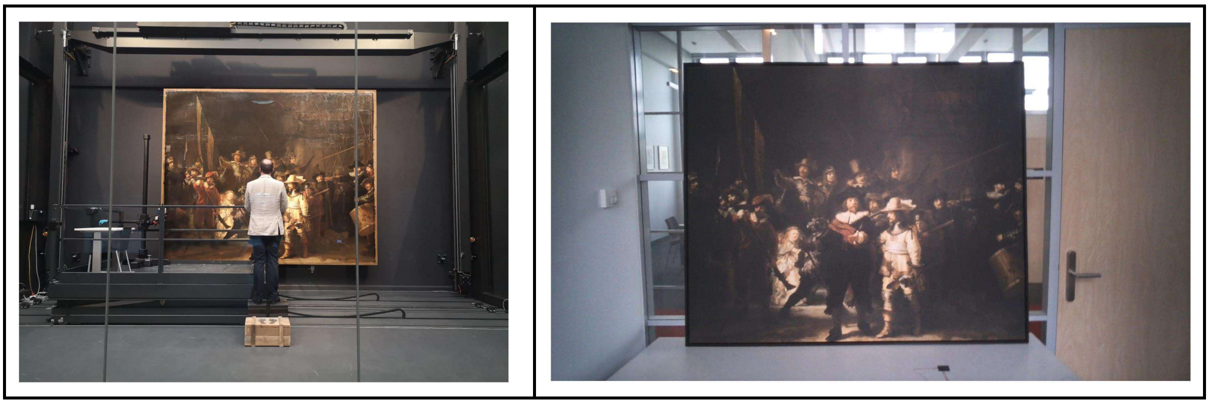

Figure 6. Experimental setup for *The Night Watch* study in a glass chamber in the Rijksmuseum (left) and a replica in an office environment (right) (De Winter et al., 2022).

4.1. Methods

The supplementary data for *The Night Watch* experiment (De Winter et al., 2022) contains two transcript files in Dutch, one for participants observing the real painting in the Rijksmuseum, and another for participants looking at the replica painting in the laboratory. A pairwise comparison was conducted between transcripts in ascending order, i.e., the first transcript in one file was compared to the first transcript in the other, and so on. Since there were a total of 21 participant transcripts in the recordings taken at the Rijksmuseum, and 26 for the replica, the new analysis included only the first 21 comparisons.

We used the online chat window, instead of the API used in Studies 1 and 2. Each pair of participants was compared in a separate chat window, using the following prompts: "*Consider the following transcripts:*", followed by a transcript related to the original painting; the next prompt was "*How does it differ from the following transcripts, in essence?*" which was then followed by a transcript from the replica painting. For each comparison, ChatGPT automatically provided a short summary in English. Next, the following prompt was given: "*What would you regard as the key difference between both transcripts?*".

To arrive at the final output, a hierarchical approach was again adopted. Specifically, once all 21 comparisons were performed, the key difference outputs were grouped into batches of 5, 5, 5, and 6. Each batch was summarized into a 200-word paragraph using the prompt "*Please provide a 200 word summary for the following text:*". The four resulting summaries were then combined and condensed into a general 400-word summary using the prompt "*Please provide a 400 word summary for the following text:*".

4.2. Results

The text below shows the final 400-word summary that was generated by ChatGPT, highlighting the differences between what participants said while observing *The Night Watch* in the museum and its replica in the laboratory:

"*The text describes the differences between two transcripts of two people's observations of a painting. The first transcript is more emotive, focusing on the personal experience of looking at the painting, showing awe and amazement, trying to understand the meaning, and expressing it as a once in a lifetime experience. The person in this transcript is deeply moved by the painting and finds it fascinating, providing a detailed description of the painting and its intricate details, dynamic lighting and highlights on the faces, and expressing its beauty and realism.*

The second transcript is more analytical, providing a more general and surface-level description of the painting, focusing primarily on the activity of the figures and their presence of weapons and musical instruments. It also notes the lighting, but makes less interpretive statements and has less focus on the individual figures and their meaning in the scene. Additionally, the second speaker seems to have a better knowledge of the painting and its historical context, including information on its restoration and the identification of specific elements such as a dog, chicken or dove, and the presence of Rembrandt himself in the painting.

The main difference between the two transcripts is the tone, focus, and level of detail provided by the speakers. The first transcript is more emotional and expresses awe and wonder towards the artwork, while the second transcript is more analytical, providing objective observations and less emotive language. Additionally, the second speaker seems to be more knowledgeable about the painting and its historical context, but less engaged with the painting, providing a more general and surface-level description of the painting."

The generated summary above indicates that ChatGPT captured how the participants looking at the real painting were 'deeply moved by the painting', and it commented on the 'dynamic lighting and highlights on the faces'. Moreover, the participants who viewed the real painting were deemed to be 'more emotional' and expressed 'awe and wonder towards the artwork', in contrast to the participants who viewed the replica, which was described as being 'more analytical, providing objective observations and less emotive language'.

5. Discussion

We examined whether ChatGPT can automatically extract valid patterns from text data. First, we examined whether a sentiment analysis of textbox responses by ChatGPT would yield sentiment scores that correlated strongly with quantitative rating-scale data. Secondly, we examined whether raw interview data, without preprocessing, could be meaningfully summarized by ChatGPT. Thirdly, we examined whether ChatGPT is powerful enough to detect differences between think-aloud transcripts related to two different experimental conditions, namely a real painting versus a replica.

For the questionnaire textbox responses, ChatGPT was highly capable of assessing the sentiment, as demonstrated by the strong correlation (r = 0.99) between the ChatGPT sentiment scores and previously computed scores obtained through rating scales. Manual quantification of sentiment for 992 entries per interface would have been challenging, as entries varied significantly in style and sometimes contained gibberish or non-English text. ChatGPT, on the other hand, was able to interpret this data and arrive at an overall score in about one minute time.

Additionally, ChatGPT was found a useful tool for summarizing interview data from human subjects. An additional benefit was that the interview data was transcribed automatically, significantly reducing the time required compared to manual transcription. Furthermore, analyzing interviews can be challenging, particularly when dealing with large amounts of data, and subjective biases of the researcher may potentially influence the interpretation of the responses (Hewitt, 2007; Morse, 2015). ChatGPT was able to address these challenges by utilizing a transparent and replicable methodology for processing interview results.

Finally, ChatGPT was found capable of highlighting subtle differences in what participants said when observing *The Night Watch* in the Rijksmuseum compared to a replica in an office environment. These results agree with the manual analysis of the think-aloud method, which revealed that the participants uttered more words related to admiration for the real *Night Watch* than for the replica. This also adds to the ongoing discussion surrounding the effects of a painting in the context of the museum (Krukar & Dalton, 2020; Pelowski et al., 2017; Specker et al., 2017) and in a laboratory setting (DiPaola et al., 2013). However, unlike the manual analysis performed by De Winter et al. (2022), which provided keywords grouped by theme, ChatGPT was able to generate a paragraph that highlighted striking differences between the major themes identified in the two settings by just submitting transcripts without further context.

It is important to note that if a single human researcher were to have generated the text output for either Study 1, 2, or 3, it may not be considered a trustworthy scientific outcome due to the limitations of human cognition, such as confirmation bias. For example, it is possible that a human researcher, especially after having invested substantial effort into conducting the experiments, may expect or hope that certain differences between the two settings arise based on preconceptions or conflicts of interest. The fact that ChatGPT, an automatic tool, generated the text removes much of this doubt and instills a sense of confidence.

It is also important to note that the use of ChatGPT resulted in a significant reduction in time compared to manually analyzing the questionnaire or think-aloud transcript data. It took the experimenter numerous hours of work to filter, sort, and label the data manually, while using the method presented in this paper only required sorting the data and copying it into the ChatGPT interface for Study 3, while for Study 1 and 2, our script produced outputs in about 1 minute and 25 minutes, respectively. Furthermore, although it is commonly advised that ChatGPT and similar large language models require prompts that provide suitable context (e.g., Jalil et al., 2023; White et al., 2023), we found that very brief prompts were sufficient. We also found that the randomness ('temperature') parameter should be set to 0 for accurate results.

ChatGPT's outputs have benefits like replicability, but they also have limitations due to potential biases. These biases arise from the fact that ChatGPT models have been trained on human-generated text and human-in-the-loop reinforcement learning (OpenAI, 2022, 2023). ChatGPT outputs could potentially introduce biases toward political leanings (Hartmann et al., 2023; McGee, 2023) or other types of biases. However, these biases are unlikely to have affected the results because ChatGPT was used to provide numeric scores or summaries of texts and not to generate new ideas.

6. Conclusion

In this paper, we tested ChatGPT's validity as a tool for analyzing text data by using it to analyze text entries from an online questionnaire, interview transcripts, and transcripts from a think-aloud study. We employed a hierarchical approach because ChatGPT has a limit to the amount of tokens it accepts. Specifically, we provided batches of text for it to summarize before producing a final score or summary. The results showed that ChatGPT produced sentiment scores that correlated highly with quantitative metrics, produced meaningful summaries of interviews, and highlighted subtleties from think-aloud data. This significantly reduced the time needed for data analysis and provided valuable insights into the experimental conditions. ChatGPT is a valid and robust tool for qualitative data analysis in HCI research.

References

Alba, D. (2022). OpenAI chatbot spits out biased musings, despite guardrail. https://www.bloomberg.com/news/newsletters/2022-12-08/chatgpt-open-ai-s-chatbot-is-spitting-out-biased-sexist-results

Bommarito, M., II, & Katz, D. M. (2022). *GPT takes the bar exam.* arXiv. https://arxiv.org/abs/2212.14402

Borji, A. (2023). *A categorical archive of ChatGPT failures.* arXiv. https://doi.org/10.48550/arXiv.2302.03494

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Chakrabarti, P., & Frye, M. (2017). A mixed-methods framework for analyzing text data: Integrating computational techniques with qualitative methods in demography. *Demographic Research*, *37*, 1351–1382. https://doi.org/10.4054/DemRes.2017.37.42

Clemmensen, T., & Roese, K. (2010). An overview of a decade of journal publications about culture and human-computer interaction (HCI). In D. Katre, R. Orngreen, P. Yammiyavar, & T. Clemmensen (Eds.), *Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts. HWID 2009* (pp. 98–112). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-11762-6_9.

De Carvalho, P., & Fabiano, A. (2021). Thematic analysis for interactive systems design: A practical exercise. *Proceedings of 19th European Conference on Computer-Supported Cooperative Work. European Society for Socially Embedded Technologies*, Zürich, Switzerland. https://doi.org/10.18420/ecscw2021_wsmc06

De Winter, J. C. F. (2022). Can ChatGPT pass high school exams on English language comprehension? https://www.researchgate.net/publication/366659237_Can_ChatGPT_pass_high_school_exams_on_English_Language_Comprehension

De Winter, J. C. F., Dodou, D., & Tabone, W. (2022). How do people distribute their attention while observing The Night Watch? *Perception*, *51*, 763–788. https://doi.org/10.1177/03010066221122697

DiPaola, S., Riebe, C., & Enns, J. T. (2013). Following the masters: Portrait viewing and appreciation is guided by selective detail. *Perception*, *42*, 608–630. https://doi.org/10.1068/p7463

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). *Mathematical capabilities of ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2301.13867

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers.* bioRxiv. https://doi.org/10.1101/2022.12.23.521610

Gerlach, J. H., & Kuo, F.-Y. (1991). Understanding human-computer interaction for information systems design. *MIS Quarterly*, *15*, 527–549. https://doi.org/10.2307/249456

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2022). *How does ChatGPT perform on the medical licensing exams? the implications of large language models for medical education and knowledge assessment.* medRxiv. https://doi.org/10.1101/2022.12.23.22283901

González-Padilla, D. A. (2022). Concerns about the potential risks of artificial intelligence in manuscript writing. *Journal of Urology*. https://doi.org/10.1097/JU.0000000000003131

Gubrium, J. F., & Holstein, J. A. (2001). *Handbook of interview research: Context and method.* Thousand Oaks, CA: Sage Publications. https://doi.org/10.4135/9781412973588

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation.* arXiv. https://doi.org/10.48550/arXiv.2301.01768

Heikoop, D. D., De Winter, J. C. F., Van Arem, B., & Stanton, N. A. (2018). Effects of mental demands on situation awareness during platooning: A driving simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58*, 193–209. https://doi.org/10.1016/j.trf.2018.04.015

Hewitt, J. (2007). Ethical components of researcher—researched relationships in qualitative interviewing. *Qualitative Health Research*, *17*, 1149–1159. https://doi.org/10.1177/1049732307308305

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., & Lam, W. (2023). *ChatGPT and software testing education: Promises & perils.* arXiv. https://doi.org/10.48550/arXiv.2302.03287

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, *42*, 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Kirmani, A. R. (2023). Artificial Intelligence-Enabled Science Poetry. *ACS Energy Letters*, *8*, 574–576. https://doi.org/10.1021/acsenergylett.2c02758

Kitto, K., Manly, C. A., Ferguson, R., & Poquet, O. (2023). Towards more replicable content analysis for learning analytics. *Proceedings of Learning Analytics and Knowledge 2023*, Arlington, TX. https://doi.org/10.1145/3576050.3576096

Kjeldskov, J., & Skov, M. B. (2003). Creating realistic laboratory settings: comparative studies of three think-aloud usability evaluations of a mobile system. In M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Human-Computer Interaction (INTERACT'03)* (pp. 663-670). Amsterdam, The Netherlands: IOS Press.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: Sage Publications.

Krukar, J., & Dalton, R. C. (2020). How the visitors' cognitive engagement is driven (but not dictated) by the visibility and co-visibility of art exhibits. *Frontiers in Psychology*, *11*, 350. https://doi.org/10.3389/fpsyg.2020.00350

Maraj, C. S., Martinez, S. G., Badillo-Urquiola, K. A., Stevens, J. A., & Maxwell, D. B. (2016). Preliminary review of a virtual world usability questionnaire. In S. Lackey & R. Shumaker (Eds.), *Virtual, Augmented and Mixed Reality: 8th International Conference* (pp. 35–46). Cham: Springer. https://doi.org/10.1007/978-3-319-39907-2_4

McGee, R. W. (2023). Is Chat GPT biased against conservatives? An empirical study. https://doi.org/10.2139/ssrn.4359405

Morse, J. M. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research*, *25*, 1212–1222. https://doi.org/10.1177/1049732315588501

OpenAI. (2022). Introducting ChatGPT. https://openai.com/blog/chatgpt

OpenAI. (2023). How should AI systems behave, and who should decide? https://openai.com/blog/how-should-ai-systems-behave

Otter.ai. (2023). Otter.ai - Voice meeting notes & real-time transcription. https://otter.ai

Patel, S. B., & Lam, K. (2023). ChatGPT: the future of discharge summaries? *The Lancet Digital Health*, *5*, E107–E108. https://doi.org/10.1016/S2589-7500(23)00021-3

Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator.* https://doi.org/10.1177/10776958221149577

Pelowski, M., Forster, M., Tinio, P. P. L., Scholl, M., & Leder, H. (2017). Beyond the lab: An examination of key factors influencing interaction with 'real' and museum-based art. *Psychology of Aesthetics, Creativity, and the Arts*, *11*, 245–264. https://doi.org/10.1037/aca0000141

Roberts, K., Dowell, A., & Nie, J.-B. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Medical Research Methodology*, *19*, 66. https://doi.org/10.1186/s12874-019-0707-y

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*. https://doi.org/10.37074/jalt.2023.6.1.9

Schelble, B. G., Flathmann, C., Musick, G., McNeese, N. J., & Freeman, G. (2022). I see you: Examining the role of spatial information in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, *6*, 374. https://doi.org/10.1145/3555099

Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). *An analysis of the automatic bug fixing performance of ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2301.08653

Specker, E., Tinio, P. P. L., & Van Elk, M. (2017). Do you see what I see? An investigation of the aesthetic experience in the laboratory and museum. *Psychology of Aesthetics, Creativity, and the Arts*, *11*, 265–275. https://doi.org/10.1037/aca0000107

Stoker-Walker, C. (2022). AI bot ChatGPT writes smart essays — should professors worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7

Tabone, W., Happee, R., García, J., Lee, Y. M., Lupetti, M. L., Merat, N., & de Winter, J. (2023a). Augmented reality interfaces for pedestrian-vehicle interactions: An online study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *94*, 170–189. https://doi.org/10.1016/j.trf.2023.02.005

Tabone, W., Happee, R., Yang, Y., Sadraei, E., García, J., Lee, Y. M., Merat, N., & De Winter, J. (2023b). Augmented reality interfaces for pedestrian-vehicle interactions: A naturalistic simulator study. Manuscript in preparation.

Tate, T. P., Doroudi, S., Ritchie, D., & Xu, Y. (2023). *Educational research and AI-generated writing: Confronting the coming tsunami.* EdArXiv. https://doi.org/10.35542/osf.io/4mec3

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*, 313. https://doi.org/10.1126/science.adg7879

Van Dis, E. A. M., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*, 224–226. https://doi.org/10.1038/d41586-023-00288-7

Vincent, J. (2022). AI-generated answers temporarily banned on coding Q&A site Stack Overflow. https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2302.11382

Zhang, X., & Simeone, A. L. (2022). Using the think aloud protocol in an immersive virtual reality evaluation of a virtual twin. *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, Online CA. https://doi.org/10.1145/3565970.3567706

Zhao, T., & McDonald, S. (2010). Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, 581–590, Reykjavik, Iceland. https://doi.org/10.1145/1868914.1868979

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). *Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT.* arXiv. https://doi.org/10.48550/arXiv.2302.10198