ArtBot: An Exploration into AI's Potential for Guiding Art Analysis

Thomas Serban von Davier
Department of Computer Science
University of Oxford
Oxford, United Kingdom
thomas.von.davier@cs.ox.ac.uk

Aaron John Henry Larsen
Faculty of History
University of Oxford
Oxford, United Kingdom
aaron.larsen@st-hildas.ox.ac.uk

Max Van Kleek
Department of Computer Science
University of Oxford
Oxford, Oxfordshire, United Kingdom
max.van.kleek@cs.ox.ac.uk

Nigel Shadbolt
Department of Computer Science
University of Oxford
Oxford, United Kingdom
nigel.shadbolt@cs.ox.ac.uk

Abstract

Art analysis, the process of reasoning through an artwork, cultivates critical thinking, empathy, and cultural awareness. However, historically, art analysis has been inaccessible to individuals outside of cultural institutions and, more recently, social media platforms, making individuals passive consumers. Therefore, we developed ArtBot, a Socratic large language model (LLM) art companion designed to guide users through artwork analysis. We evaluated this prototype through a within-subjects lab study (n=13), comparing it to conditions inspired by digital museum collection sites and social media. Our findings reveal statistically significant differences between the conditions across several metrics, including self-reported understanding, the use of complex vocabulary, and writing proficiency. Post-hoc tests revealed that ArtBot and digital collections performed comparably and slightly better than social media, but not significantly. These results suggest that AI tools can support deeper engagement with art in digital spaces while laying the groundwork for iterative testing.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

artificial intelligence, art, prototype, within-subjects experiment, user experience design, artifact

ACM Reference Format:
Thomas Serban von Davier, Aaron John Henry Larsen, Max Van Kleek, and Nigel Shadbolt. 2025. ArtBot: An Exploration into AI's Potential for Guiding Art Analysis. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706599.3720181

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720181

1 Introduction

Engaging with art is more than just observation; it is an exercise in "reasoned perception" that sharpens the mind in various ways [42]. By applying logic and reflection to artistic experiences, individuals explore the relationship between the work and their interpretation of it [18]. This process unlocks a range of vital skills—creativity, cultural awareness, empathy, and critical thinking—that extend far beyond the art world [18, 24, 42, 54]. However, despite these benefits, access to art analysis remains limited. Social theory historically suggested that institutions and cultural elites guard this knowledge, keeping it out of reach for many [1, 7, 25]. Recently, scholars argue social media platforms act as "infomediaries" and "gatekeepers," pushing users into passive consumption of algorithm-driven content [29, 30]. Human-computer interaction (HCI) researchers are calling for a shift [43], advocating for artificial intelligence (AI) tools that empower users to critically engage with digital content [39, 53]. We argue that art analysis is not just a skill but a critical foundation for developing essential thinking skills in our increasingly digital world.

We introduce ArtBot, an art companion built on a large language model (LLM) designed to act as a Socratic opponent, an agent designed to challenge the user with questions while observing artworks. ArtBot is a concrete prototype for exploring the potential of a new algorithmic experience (AX) [2]. Internally, ArtBot uses Retrieval Augmented Generation (RAG) upon a custom knowledge base to provide depth and completeness about art pieces being discussed, comprising of a combination of open-source information about these works, curator-provided information, and additional open data collected from auction house records [51].

To understand the potential and challenges for ArtBot as a Socratic opponent for art appreciation, we conducted an exploratory study with the prototype through a within-subjects experiment (n=13). The experiment had participants view nine artworks in three randomized conditions that replicated three art experiences. One condition replicates a digital collection page containing art and some wall text written by a curator or art historian. Another replicates a social media post with an image and a basic label underneath. The third condition is ArtBot, where the AI accompanies the image and label. Our research questions are as follows:

- **RQ1** - Could a Socratic opponent for digital art experiences help art observers experience art with a critical perspective?
- **RQ2** - What are the design challenges and opportunities in building AI-driven Socratic opponents for art experiences?

We found statistically significant differences between the conditions on three primary metrics: self-reported understanding, the use of complex vocabulary, and writing proficiency. Post-hoc tests revealed that the prototype performed at the same level as a digital collection experience in some of these metrics, both of which outperformed the social media experience.

Based on these findings, we discuss how tools like these could be expanded to better collaborate with users as they review creative work and develop their perspectives.

2 Related Work: Art Analysis for the People

Art analysis is a skill that highly benefits individuals [1]. Nonetheless, outside formal art education, the skillset is not accessible to everyone. Studies have shown that factors like socioeconomic status and education level are related to one's likelihood of conversing about art, the type of conversation one will have, and the type of art one will likely discuss [25]. The societal divide around art is not a new phenomenon; many scholars refer back to Bourdieu's concept of "cultural intermediaries" when discussing how fragmented access to art analysis and education is among different groups [7]. The term referred traditionally to institutions like universities and museums that have been the gateway to cultural reflection as cultural exposure moved from these physical institutions to digital spaces, as did the evaluation of art. While more pieces were accessible to audiences, the language and depth of knowledge were arguably lost [43]—eventually, click-bait captions and opaque recommendations replaced expert wall text and curated exhibits.

Morris has labeled this transition as creating cultural "infomediaries" [30]. The idea is that recommender systems on social media platforms dictate what we see and consume. Manovich agrees with this argument by describing how social media impacts the public's aesthetic values and tastes as a universal curator [26]. While social media is built to collect and monetize user data, scholars, even among the HCI community, argue it still plays the role of a curator dictating how and what we see [14, 21].

In both cases, the ability of any individual to gain adequate access to art information and analyze the content they are seeing is limited. To combat this lack of access, many museums have attempted to develop open-access digital collections [45] or establish their presence on social media [22]. However, not all systems are fully digitized; many are still catching up with recording all the information in their massive collections, or they struggle to achieve the appropriate tone of voice for social media. Due to this existing lack of access, we argue that ArtBot has the potential to democratize high-quality art information to wider audiences.

In the study, we compare ArtBot's performance to that of a digital collection and a social media example. Based on the results, ArtBot may become more accessible than a traditional institutional experience without sacrificing quality and the benefits of art analysis, like the shallower social media condition.

3 Developing ArtBot

ArtBot is our approach to democratizing art analysis through a digital experience for a wide range of users. Its development followed four stages: gathering design requirements from literature, accessing quality art data, testing customizable LLMs, and finding open-access artwork.

Starting with design requirements, we turn towards recent research into literature on algorithmic experiences of art recommendation. Recent studies have argued that art experiences need "active, beneficial partnerships rather than one-sided content pipelines" [56]. This motivated us to consider how a large language model could be paired with a displayed artwork, allowing users to interact and discuss the art directly with the algorithmic system. A second inspiration area for the design was the speculative design approach of Slow Technology [20, 33]. The original paper argues that researchers can prompt reflection in users by slowing down experiences that tend to be accelerated by technology. As we are looking to improve art analysis by challenging the speed and shallow interactions of social media, the design philosophy of Slow Technology was incorporated into our development of ArtBot. Finally, we reviewed recent HCI papers that featured the development of experimental prototypes [38, 46]. Their insights into the degree of fidelity and delivery system of the prototypes also informed our approach.

With our design approach solidified, we turned toward the system's back end. As we set out to build a customized large language model, one of the challenges became equipping it with knowledge specifically about artworks. The answer was to build the model as a retrieval-augmented generative (RAG) model. These models rely on a unique dataset the LLM can encode and then parse to improve and specialize its answers. However, that requires a specialized dataset of art data, and as stated in Section 2, digital collections are often incomplete. Therefore, we turned to the open-source dataset, *AppraiSet* [51]. This dataset was built on art auction data, which included information about artists and artworks through provenance and condition reports. HCI research has explored the dataset to improve art metadata exploration [52].

We decided to utilize local LLMs for the rapid prototyping needed to develop ArtBot. The basic model details came from open-source tools like Ollama [3], which allowed us to download various local models for early experimentation and design. Having a local model meant we could keep some level of control over it as we worked to fit it within the overall algorithmic experience of the prototype. Running the prototypes and eventual experiments locally meant we could account for the data protection practices required by our institution's research ethics committee. Furthermore, we could develop our model files, which could be customized for different parameters and system prompts to customize the model. The unique capability of this prototype is the ability to prompt the user. This was inspired by education and critical thinking literature [12, 37, 55]. Building the LLM to function as a Socratic opponent, an agent designed to question the user about artworks, for art education was the main task of the system prompt.

With the prototype functionality established, the final step was displaying art objects. We contacted The Ashmolean Museum,

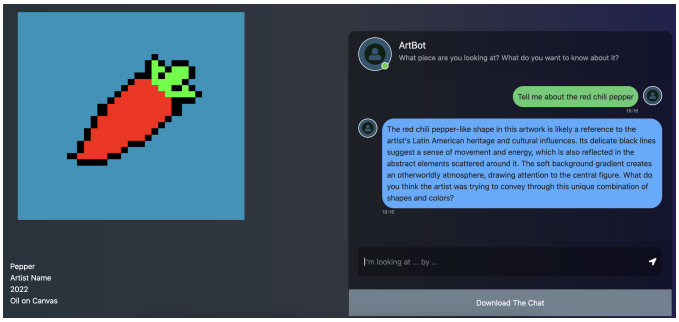


Figure 1: This is an image of the ArtBot experience. An artwork is presented to the user on the left side, while the AI interaction is available on the right side.

which is free to the public and has an open digital collection. Working within their terms of service, we selected nine paintings randomly from their heavily featured Western art section, ranging across different art styles spanning 200 years. This was done to avoid heavily featuring any single artist or style, which may sway opinions. We acknowledge that only using paintings from Western Artists heavily skews the prototypes towards a particular perspective. The decision was made based on the prototype's image quality and presentation capabilities. While the museum features artwork from all over the world, the pieces from other cultures were often artifacts better viewed in three dimensions, which is not a capability the prototype was built for. Therefore, paintings were the best option for displaying the artwork, and that limited us to their extensive collection of Western art paintings.

3.1 The Experiment Ready Version

The experiment-ready version of ArtBot was built on a Llama 3.1 model [28] with specific art support from the *AppraiSet* art dataset [51]. The nine artworks and their exhibit wall text were added to the dataset. To ensure the dataset also had information about the visual contents of the image, we had GPT4 visually analyze the artworks before outputting a paragraph description of the work. A custom system prompt managed all of this (see Appendix A). With ArtBot working, the final step was to package it in a local web app for testing. See Figure 1 for a standard view of ArtBot. The artwork is always displayed on the left, with the caption below it, offering the user a starting point to begin the conversation. We have placed questions and example text around the chat window and within the chatbox to guide the user in initiating the conversation. Naturally, due to the Socratic method implemented in the system prompt, once the first message is sent, the ArtBot will take over, prompting the user with its questions about the work. The user can engage with the conversation as long as they wish and exit by clicking away or using a Llama 3.1 stop word. For testing purposes, we also included a button that will allow us to download the contents of the chat locally for diagnostics and further analysis. Readers can find the base code in the Supplementary Materials.

4 Experimental Methods

4.1 Participant Recruitment

Due to the controlled nature of the prototype, participation was done in person in a controlled academic space (study rooms, meeting rooms, and classrooms). Therefore, participant recruitment was also done in person on a university campus through advertising posters, in-person flyering and advertising, and direct messages sent across the researchers' networks. Informed consent was collected before conducting any research or data collection. All of the user research was approved by our institutional ethic review (approval code: CS_C1A_24_019)

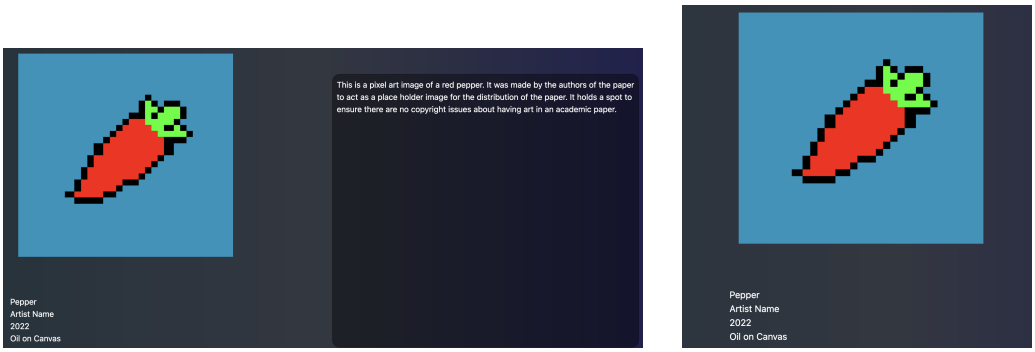
In the end, 13 participants were recruited, mostly graduate students aged 18-44. Participants represented six global regions (Middle East, Europe, Africa, Oceania, South America, and South Asia). The gender breakdown was approximately 15% non-binary, 38% male, and 47% female. Regarding their experience with art and art museums, participants reported going to museums approximately 2-12 times in the last year. Two participants reported having gone fewer than twice in the last year, and one reported going over 25 times within the last year. Their average self-reported knowledge of Art is 3.31 on a scale from 1-7, where 1 is a complete novice, and 7 is a student of Art. Similarly, their comfort with LLMs and chatbots was, on average, 3.46 on a scale from 1-7, where 1 had never used LLMs, and 7 had used them multiple times. See Table 1 for an overview. All participants considered themselves fluent in English, if not native speakers.

4.2 Study Design

We conducted a within-subjects experiment comparing ArtBot to two other conditions reflecting current art experiences. This design controls for participant variance, particularly given our highly educated participants with diverse experience in art and LLMs. Each participant interacted with all three conditions, each repeated three times to ensure reliability. We randomly assigned nine images per participant to prevent fatigue and favoritism to specific artworks (three per condition). Inspired by recent HCI studies [46], our approach integrates an experimental prototype with a formal study. The order of conditions and image assignments was fully randomized to mitigate ordering effects [41] and reduce variability introduced by different artworks. See Table 2 for a breakdown of the study design.

Digital Collection. The digital collection condition aims to replicate a page on a museum website. With greater digital resources available, digital collection pages have become popular offerings on museum websites [5, 8]. These digital collections are also often the subject of redesigns by researchers [6, 19, 27], making them a well-established user experience against which ArtBot can be compared. The participants are presented with an image with a label including the title, year, artist name, and materials. The image and label are supported by a plaque of "wall text" written by a curator or art historian. See Figure 2a.

Social Media. The social media condition replicates a social media post. In this case, the image is central to the digital screen and accompanied by a short label. The label, again, is made of the title, year, artist name, and materials. See Figure 2b.



(a) **(b)**

Figure 2: This figure contains two images of the experimental conditions for the digital collection (a) and social media post (b). Each show an image presented to the user with varying degrees of

ArtBot. The final condition is our prototype. The image and label are kept the same, but instead of having a plaque with wall text, the participants are presented with the interaction area where they can chat with the LLM about the piece.

Following each condition, the participants are presented with a Likert question asking how much they understood about the artwork from 1-7. Then, they were asked one larger text question, asking them to explain their position. The combination of a short reaction question followed by a question asking for justification is pulled from education curricula [16, 31, 34]. All 13 participants answered these questions, but one participant's form was deleted due to an issue with Microsoft Forms.

4.3 Data Analysis

From each participant, we received nine Likert responses and nine essay responses, one of each for every randomized image and condition. For the Likert scale responses, we used a χ^2-test for independence [41] comparing the categorical values of the scale responses with each test condition. If the test returns a statistically significant p-value < 0.05, we conducted posthoc Bonferroni adjustments [41] to see which, if any, combination of variables was statistically significant. We used the stats module from the scipy Python library [50] for these calculations.

We used two analysis methods for the essay responses: computational text analysis and human grading. The text analysis assessed quality, complexity, and readability, while the human grader evaluated understanding and artistic concept mastery based on curriculum standards. The following paragraphs detail each approach.

The computational text analysis method comes from the textstat Python library [4] and allowed us to gain measures on the word length, number of complex words, and predicted grade level. The grade level prediction score combined multiple grade level predictors to output an estimated year range (9th to 10th grade), which we marked as a value of 9.5. A score like this would correspond to a student in the first year of secondary school. The calculation combines established readability scores that have been used and built upon over the last 80 years [9, 10, 13, 15, 36]. We treated this as a continuous variable as these measures correspond with academic year lengths from 0.5 - 23.5. Once all of the text analysis metrics were calculated, we used the statsmodels Python library [40] to perform repeated measures ANOVA tests to see if there was a statistically significant difference between the means of the three conditions. If the test revealed a statistically significant difference, we performed post-hoc Tukey method tests [41] to see which of the three conditions differed from each other.

The second author has formal education training and professional experience as an educator in multiple nations. Their specific work has included museum educational outreach aimed at teaching school children of various ages how to examine and describe the artworks they see. They served as the human assessor for the text responses and were kept blind from the results of the other analyses and which responses came from the conditions. The second author developed a rubric (see Table 3) to evaluate the essay responses. The rubric was built on the foundations of the British Columbia Curriculum, first fully implemented in 2019 [34]. Following the Provincial Proficiency Scale, the responses were assessed along the criteria of Emerging, Developing, Proficient, and Extending [35]. Similar arts education policies are in place in the USA [31] and United Kingdom [16]. The Arts Education curriculum for Secondary Students highlights the significance of applied knowledge, personal connections, and clear communication when discussing and reflecting on art. The curriculum determined the qualities the rubric assessed. The use of ArtBot replicates the technique of dialogic teaching, a pedagogy that enshrines learning within the framework of a conversation [44]. While the algorithmic assessment of the responses was purely focused on the quantifiable data, the responses following a dialogue with the AI present similar levels of development and growth as a traditional dialogic classroom lesson: "...it would be appropriate to base the assessment of students' literacy development, at least in part, on an examination of the communicative competence they display in structured group discussions about the texts they have read." [44]. Following the development of the rubric, the assessor also evaluated all of the text responses, providing them with scores ranging between 4 and 16.

5 Results

5.1 Likert Responses

Likert scale ratings of understanding an artwork are categorical comparisons across the randomized research conditions. Therefore, our null hypothesis is that the reaction to the artwork is independent of the conditions.

In their self-reported understanding of the artwork, the χ^2-test returned a statistically significant result (statistic=21.41, p-value=0.044). This means we could reject the null hypothesis and state that there is a relationship between self-reported understanding of the artwork and the conditions. The proportional distribution of responses across the conditions can be seen in Table 4. There appear to be more instances of greater understanding among the Digital Collection and AI conditions with more instances of lesser understanding in the Social Media condition. While differences are based on the χ^2-test, the Bonferroni adjustment results show we cannot say which differences are significant. In this instance, both the Digital Collection and AI conditions record higher understanding ratings, while social media predominantly has lower levels.

5.2 Text Responses

In evaluating the text responses computationally, our repeated measures, ANOVA, have the null hypothesis that there will be no difference in performance across the three conditions when we examine the length of text, the number of complex words, and the grade level of the writing. We could not reject the null hypothesis for the length of text; the number of words written did not alter statistically based on the condition.

We did receive statistically significant ANOVA calculations based on the number of complex words across the conditions and the grade-level calculation. For the number of complex words (F = 3.78, p-value = 0.038), we could reject the null hypothesis and state that there is a difference based on the condition. As seen in Figure 3a, the variance in complex word counts for participants in the Digital Collection and AI condition was relatively high, even though their averages were above that of the Social Media condition. Due to this large variance, the post-hoc Tukey test could not determine which of the conditions were statistically different from each other. In the grade-level analysis (F=6.38, p-value = 0.006), we can reject the null hypothesis and state that grade level differs based on the experimental conditions. In Figure 3b, the variance between the Digital Collection and AI conditions overlap while the Social Media condition is lower than either one; this is further reinforced by the posthoc Tukey measurement that identifies statistically significant differences between the Digital Collection-Social Media conditions and the AI-Social Media conditions. The post-hoc test did not find a difference between the Digital Collection-AI conditions, indicating they are statistically comparable.

After the grading was completed, the scores associated with each text response delivered by the human assessor were also gathered for repeated measures of ANOVA. With the null hypothesis set that there would be no difference in performance across the three conditions, the test returned a non-significant result (F = 1.53, p-value = 0.239) meaning we cannot reject the null hypothesis. When the averages and variance were plotted in Figure 3c, there is evidence that the variety in response quality across users in the conditions meant the differences between conditions were insignificant.

6 Discussion

In this work, we present the results of our exploratory study comparing ArtBot to digital collections and social media conditions. Based on the quantitative analysis, the prototype outperformed the social media condition on self-reported understanding, usage of complex language, and writing proficiency.

Based on the findings, we discuss the ability of ArtBot to expand access to art analysis (Section 6.1). We reflect on whether ArtBot delivers on its ability to be more accessible while still upholding the quality of information expected of an art companion. We also connect our findings back to ongoing discussions in the field of HCI on the relationship between user satisfaction and system performance.

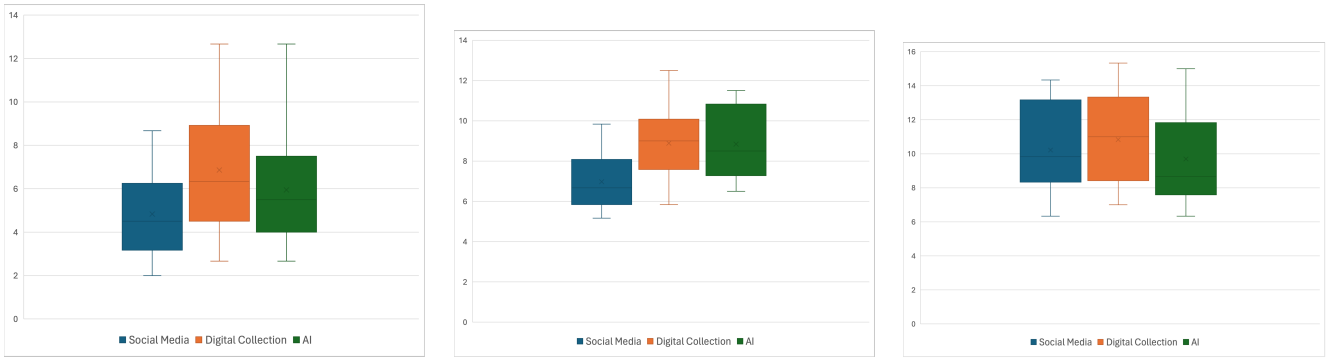
6.1 Did it Work?

ArtBot was successful on two fronts: how it performed compared to the other conditions and how our participants received it. We measured performance quantitatively across all the conditions in a randomized within-subjects study. Based on the realities of art analysis access for individuals outside of institutions (outlined in Section 2), we set out to see if ArtBot can digitally deliver art experiences without sacrificing the depth of information (**RQ1**). In other words, we want to see ArtBot outperform social media while performing the same or better than the digital collection and social media experiences.

Our results confirm this hypothesis. Our findings identified significant differences between the conditions regarding self-reported understanding of the artworks viewed and the complexity of their essay response. We found that the AI and Digital Collection conditions had a higher percentage of users reporting a deeper understanding than the social media condition. Simultaneously, ArtBot and the Digital Collection outperformed the social media condition in the text responses. This suggests that ArtBot offers a unique digital experience that delivers the same level of information as the Digital Collection condition. By comparing our findings to the literature cited in Section 2, we see how ArtBot compares to the cultural intermediaries [7] and infomediaries [30]. As these cultural institutions are the current benchmark for critical engagement with art, we can answer **RQ1** that ArtBot and Socratic systems can deliver critical art experiences for audiences.

ArtBot and digital collection experiences outperformed the social media condition, supporting Morris's and other HCI scholars' warnings about the rise of algorithmically powered recommendation systems and their influence on our artistic and cultural experiences [14, 21, 26, 30]. We can provide early evidence urging against the reliance on infomediaries to decide what aspects of art and culture we encounter.

RQ2 considered the challenges and opportunities of ArtBot and Socratic LLM systems. One major challenge drawn from the findings is balancing the usage of ArtBot with digital collections. To consider this decision we must explore which experience provides greater user satisfaction. User satisfaction has been a qualitative question common in HCI literature [23]. Notably, in 1994, Nielsen et al. and Gatian explored the power of user satisfaction and the instances



(a) (b) (c)

Figure 3: The three figures depict the box plots for analyzing the essay question of our participants. Each one depicts the three conditions: social media (blue), digital collection (orange), and AI (green). The order of the measurement is difficult words (a), grade level (b), and human assessor (c). Full-size images can be seen in Appendix B.

arguing that high user satisfaction often relates to positive outcomes when using a system or device [17, 32]. Beyond this LBW paper, we will explore our qualitative data relating to ArtBot to find answers to user satisfaction compared to the other conditions.

The major opportunity, in response to **RQ2**, resides in the accessibility to art analysis and arts education. The technological application of language models to art data allows entirely new interactions for users to explore. Therefore, we propose comparing ArtBot with other AI-powered tools like the Living Museum app [47], various LLM museum guides [48, 49], or CulturAI [11]. As all of these tools are early implementations of novel technology, it is valuable for researchers to gain an understanding of their capabilities and limitations. Our paper provides evidence that these tools can provide access to art analysis in digital environments without losing the quality expected of a museum site.

7 Limitations and Future Work

ArtBot remains an experimental system that requires further testing and development. Future work can expand the type of art presented from a more global collection to reflect the international user group, which we plan to expand beyond 13 participants. Additionally, the study only used paintings, which is only one type of medium; future work can explore other forms by building in video support. Furthermore, we only replicate one form of social media, not the widely popular short-form video social media. Again, a future iteration of this study could compare ArtBot to this type of interaction. Finally, we already have a body of qualitative data that needs to be analyzed for additional iterations on ArtBots design beyond this Late-Breaking Work.

We are looking forward to future work on ArtBot and related AI technologies. We have made the code open-source, encouraging further exploration of novel art experiences.

8 Conclusion

In this paper, we set out to deliver a novel algorithmic experience for art analysis, a skill previously inaccessible to individuals outside cultural institutions or powerful platforms. We presented the development of ArtBot, a Socratic LLM art companion, and tested it compared to digital collection and social media experiences. The findings revealed that ArtBot offers similar benefits to art analysis as a digital collection does with wall text next to an artwork. We argue that the findings support greater access to art analysis, which aids individuals with critical thinking. We present ArtBot as an early example of critical thinking support AIs that offer a novel interaction experience requiring users to engage more actively with their own cognitive processes.

Acknowledgments

We would like to acknowledge all of our participants who took the time to assist in doing this research. By taking time out of their days, their insights provided us with unique perspectives and new questions for future research. We would also like to thank the rest of our HCAI group for their support throughout this paper-writing process.

References

[1] Catharine Abell. 2012. Art: What it Is and Why it Matters. *Source: Philosophy and Phenomenological Research* 85 (2012), 671–691. Issue 3. https://about.jstor.org/terms

[2] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018). https://doi.org/10.1145/3173574.3173860

[3] Multiple Authors. 2023. Ollama. https://github.com/ollama/ollama.

[4] Shivam Bansal and Chaitanya Aggarwal. 2014. textstats. https://github.com/textstat/textstat.

[5] Enrico Bertacchini and Federico Morando. 2011. The Future of Museums in the Digital Age: New Models of Access and Use of Digital Collections. *International Journal of Arts Management* 15 (1 2011). Issue 2.

[6] Bernadette Biedermann. 2017. 'Virtual museums' as digital collection complexes. A museological perspective using the example of Hans-Gross-Kriminalmuseum. *Museum Management and Curatorship* 32 (5 2017), 281–297. Issue 3. https://doi.org/10.1080/09647775.2017.1322916

[7] Pierre Bourdieu. 1993. *The Field of Cultural Production.* Columbia University Press. 1–322 pages.

[8] Fiona Cameron. 2003. Digital Futures I: Museum Collections, Digital Technologies, and the Cultural Construction of Knowledge. *Curator: The Museum Journal* 46 (7 2003), 325–340. Issue 3. https://doi.org/10.1111/J.2151-6952.2003.TB00098.X

[9] J P Peter Kincaid Robert Fishburne Jr Richard L Rogers Brad S Chissom. 1975. Derivation Of New Readability Formulas (Automated Readability Index, Fog

Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. (1975). http://library.ucf.edu

[10] Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60 (4 1975), 283–284. Issue 2. https://doi.org/10.1037/H0076540

[11] Nicolas Constantinides, Argyris Constantinides, Dimitrios Koukopoulos, Christos Fidas, and Marios Belk. 2024. CulturAI: Exploring Mixed Reality Art Exhibitions with Large Language Models for Personalized Immersive Experiences. *UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (6 2024), 102–105. https://doi.org/10.1145/3631700.3664874/SUPPL_FILE/CULTURAI_SYSTEM_DEMO_UMAP_2024.MP4

[12] Martín Cáceres, Miguel Nussbaum, and Jorge Ortiz. 2020. Integrating critical thinking into the classroom: A teacher's perspective. *Thinking Skills and Creativity* 37 (9 2020), 100674. https://doi.org/10.1016/J.TSC.2020.100674

[13] Edgar Dale and Jeanne S. Chall. 1948. *A Formula for Predicting Readability*. Vol. 27. Ohio State University, Bureau of Educational Research. 11–20, 37–54 pages.

[14] Ankolika De and Zhicong Lu. 2024. #PoetsOfInstagram: Navigating The Practices And Challenges Of Novice Poets On Instagram. *ACM CHI 2024. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI'24)* 1 (2 2024). https://doi.org/10.1145/1122445.1122456

[15] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (6 1948), 221–233. Issue 3. https://doi.org/10.1037/H0057532

[16] Department for Education. 2013. National Curriculum - Art and design key stage 3. www.nationalarchives.gov.uk/doc/open-

[17] Amy W. Gatian. 1994. Is user satisfaction a valid measure of system effectiveness? *Information & Management* 26 (3 1994), 119–131. Issue 3. https://doi.org/10.1016/0378-7206(94)90036-1

[18] George Geahigan. 1996. Conceptualizing Art Criticism for Effective Practice. *Source: The Journal of Aesthetic Education* 30 (1996), 23–42. Issue 3. https://www.jstor.org/stable/3333320?seq=1&cid=pdf-

[19] S. E. Hackney and Zoe Faye Pickard. 2018. Creating Digital Collections: Museum Content and the Public. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10766 LNCS (2018), 626–631. https://doi.org/10.1007/978-3-319-78105-1_71

[20] Lars Hallnäs and Johan Redström. 2001. Slow Technology-Designing for Reflection. *Personal and Ubiquitous Computing* 5, 3 (2001), 201–212.

[21] Laura M Herman. 2021. Remixing, Seeing, and Curating: Algorithms' Influence on Human Creativity. *Creativity & Cognition* (2021). https://doi.org/10.1145/3450741.3467464

[22] Emma June Huebner. 2022. TikTok and museum education: A visual content analysis. *International Journal of Education Through Art* 18 (6 2022), 209–225. Issue 2. https://doi.org/10.1386/ETA_00095_1/CITE/REFWORKS

[23] Daniel R. Ilgen and Bruce W. Hamstra. 1972. Performance satisfaction as a function of the difference between expected and reported performance at five levels of reported performance. *Organizational Behavior and Human Performance* 7 (6 1972), 359–370. Issue 3. https://doi.org/10.1016/0030-5073(72)90022-0

[24] F. Kester. 2019. Including a new audience : A study into the perception of art of culturally inexperienced young adults. (1 2019).

[25] Omar Lizardo. 2016. Why "cultural matters" matter: Culture talk as the mobilization of cultural capital in interaction. *Poetics* 58 (10 2016), 1–17. https://doi.org/10.1016/J.POETIC.2016.09.002

[26] Lev Manovich. 2023. *Artificial Aesthetics: A Critical Guide to AI, Media and Design*. http://manovich.net/index.php/projects/artificial-aesthetics

[27] Paul F. Marty. 2011. My lost museum: User expectations and motivations for creating personal digital collections on museum websites. *Library & Information Science Research* 33 (7 2011), 211–219. Issue 3. https://doi.org/10.1016/J.LISR.2010.11.003

[28] Meta. 2024. Llama3.1. https://llama.meta.com/.

[29] Cheryl Metoyer-Duran. 1993. Information Gatekeepers. *Annual Review of Information Science and Technology (ARIST)* 28 (1993), 111–50.

[30] Jeremy Wade Morris. 2015. Curation by code: Infomediaries and the data mining of taste. *European Journal of Cultural Studies* 18 (6 2015), 446–463. Issue 4-5. https://doi.org/10.1177/1367549415577387/FORMAT/EPUB

[31] NCCAS. 2016. National Core Arts Standards: A Conceptual Framework for Arts Learning. www.nationalartsstandards.org.

[32] Jakob Nielsen and Jonathan Levy. 1994. Measuring usability: Preference vs. Performance. *Commun. ACM* 37 (1 1994), 66–75. Issue 4. https://doi.org/10.1145/175276.175282/ASSET/C9B59D25-8EAB-44A9-95B6-DC85DA6634A6/ASSETS/175276.175282.FP.PNG

[33] William Odom, Richard Banks, Abigail Durrant, David Kirk, and James Pierce. 2012. Slow Technology: Critical Reflection and Future Directions. *Proceedings of the Designing Interactive Systems Conference.* (2012), 816–827.

[34] Government of British Columbia. 2019. Arts Education K-9-Curricular Competencies Grade Exploring and creating Reasoning and reflecting Communicating and documenting Connecting and expanding K. www.curriculum.gov.bc.ca

[35] Government of British Columbia. 2019. The Provincial Proficiency Scale. https://curriculum.gov.bc.ca/sites/curriculum.gov.bc.ca/files/images/curriculum/en-proficiency-scale.jpg

[36] John O'Hayre. 1966. Gobbledygook Has Gotta Go. http://training.fws.gov/history/HistoricDocuments.html

[37] John P. Portelli. 1994. THE CHALLENGE OF TEACHING FOR CRITICAL THINKING. *McGill Journal of Education / Revue des sciences de l'éducation de McGill* 29 (4 1994). Issue 002. https://mje.mcgill.ca/article/view/8165

[38] Anna Marie Rezk, Auste Simkute, Ewa Luger, John Vines, Chris Elsden, Michael Evans, and Rhianne Jones. 2024. Agency Aspirations: Understanding Users' Preferences And Perceptions Of Their Role In Personalised News Curation. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (5 2024), 1–16. https://doi.org/10.1145/3613904.3642634

[39] Advait Sarkar, Xiaotong Xu, Neil Toronto, Ian Drosos, and Christian Poelitz. [n. d.]. When Copilot Becomes Autopilot: Generative AI's Critical Risk to Knowledge Work and a Critical Solution. https://earthweb.com/excel-users/,

[40] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[41] Howard J. Seltman. 2018. *Experimental Design and Analysis*. Carnegie Mellon University Statistics. https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

[42] Richard Siegesmund. 1998. Why Do We Teach Art Today? *Studies in Art Education* 39 (1998), 197–214. Issue 3. https://doi.org/10.1080/00393541.1998.11650024

[43] Ellen Simpson and Bryan Semaan. 2023. Rethinking Creative Labor: A Sociotechnical Examination of Creativity & Creative Work on TikTok. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (2023). https://doi.org/10.1145/3544548.3580649

[44] David Skidmore and Kyoko Murakami. 2016. *Dialogic Pedagogy: An Introduction*. Multilingual Matters, 1–16. http://ebookcentral.proquest.com/lib/bath/detail.action?docID=4614619.

[45] Jeff Steward. 2015. Harvard Art Museums API. https://github.com/harvardartmuseums/api-docs.

[46] Xiaoqing Sun, Jingyi Wang, Yan Zhou, Suhan Wang, Yixuan Li, and Xipei Ren. 2024. CO-Coffee: A Technology Probe Study to Facilitate Coffee Breaks in Open Offices. *Conference on Human Factors in Computing Systems - Proceedings* (5 2024). https://doi.org/10.1145/3613905.3651030/SUPPL_FILE/3613905.3651030-TALK-VIDEO.VTT

[47] Jonathan Talmi. 2024. Living Museum App. https://www.livingmuseum.app/

[48] Georgios Trichopoulos, Markos Konstantakis, George Caridakis, Akrivi Katifori, and Myrto Koukouli. 2023. Crafting a Museum Guide Using ChatGPT4. *Big Data and Cognitive Computing 2023, Vol. 7, Page 148* 7 (9 2023), 148. Issue 3. https://doi.org/10.3390/BDCC7030148

[49] Iva Vasic, Hans Georg Fill, Ramona Quattrini, and Roberto Pierdicca. 2024. LLM-Aided Museum Guide: Personalized Tours Based on User Preferences. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 15029 LNCS (2024), 249–262. https://doi.org/10.1007/978-3-031-71710-9_18/TABLES/2

[50] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[51] Thomas Serban von Davier, Max Van Kleek, and Nigel Shadbolt. 2022. AppraiSet. https://doi.org/10.17632/2nfvz8g27c.1

[52] Thomas Şerban Von Davier. 2023. Designing for Appreciation: How Digital Spaces Can Support Art and Culture. *Conference on Human Factors in Computing Systems - Proceedings* (4 2023). https://doi.org/10.1145/3544549.3577041

[53] Andre Ye, Jared Moore, Rose Novick, Amy X Zhang, and Paul G Allen. 2024. Language Models as Critical Thinking Tools: A Case Study of Philosophers. (4 2024). https://arxiv.org/abs/2404.04516v2

[54] Bereket A. Yilma, Chan Mi Kim, Gerald C. Cupchik, and Luis A. Leiva. 2024. Artful Path to Healing: Using Machine Learning for Visual Art Recommendation to Prevent and Reduce Post-Intensive Care Syndrome (PICS). *Conference on Human Factors in Computing Systems - Proceedings* (5 2024). https://doi.org/10.1145/3613904.3642636/SUPPL_FILE/3613904.3642636-TALK-VIDEO.VTT

[55] Lei Zhang, Hui Zhang, and Kai Wang. 2020. Media Literacy Education and Curriculum Integration: A Literature Review. *International Journal of Contemporary Education* 3 (3 2020), 55. Issue 1. https://doi.org/10.11114/IJCE.V3I1.4769

[56] Thomas Şerban von Davier, Laura M. Herman, and Caterina Moruzzi. 2024. A Machine Walks into an Exhibit: A Technical Analysis of Art Curation. *Arts 2024, Vol. 13, Page 138* 13 (8 2024), 138. Issue 5. https://doi.org/10.3390/ARTS13050138

A ArtBot System Prompt

You are an art history tutor listening and responding to someone's observations about an artwork. Use the following pieces of retrieved context to answer their questions. Limit your response to 4 sentences. End every answer with one relevant question.
Question: {question}
Context: {context}

B Supporting Charts and Figures

Table 1: An overview of our participants and their relationship to art and LLMs before the study began.

Participant	Art Knowledge (1-7)	LLM Comfort (1-7)
P1	2	4
P2	3	1
P3	2	2
P4	1	2
P5	3	3
P6	7	7
P7	6	6
P8	3	3
P9	6	2
P10	3	2
P11	3	5
P12	1	3
P13	3	3

Table 2: The three conditions of the within-subjects experiment.

Condition	Num. of Images	Situation
Digital Collection	3	Presented an image along with the official wall text. Participants could view the art and read the text before responding.
Social Media	3	Presented an image with just a caption including artist name and title, no wall text.
ArtBot (AI)	3	Presented an image with the caption, but have the interaction with the LLM before responding.

Table 3: The rubric developed to score the study participants' text responses. Built on the literature of various national curricula.

Category	Emerging (1)	Developing (2)	Proficient (3)	Extending (4)
Emotional Connection	I can identify what binary emotion this painting makes me feel.	I can identify what binary emotion this painting makes me feel, and a scale of this emotion.	I can identify how this painting makes me feel beyond a binary emotion. I can use complex language to identify one or more emotions and explain why.	I can identify my emotional connection to this piece and compare it to previous experiences or learnings from my life.
Interpretation	I cannot identify what the painting is depicting.	I can make a guess about what the painting is depicting.	I can use clues to make a guess about what the painting is depicting.	I can use clues to confidently identify my own interpretation of the painting, then explain my answer in detail.
Language	I can use simple words without sentences.	I can write in sentences using simple words to express simple ideas.	I can write multiple sentences to express my ideas.	I can write multiple sentences to express complex and abstract ideas.
Information Retention	I cannot use the information provided in my response.	I can use the information provided about the piece to make a simple observation.	I can use the information provided to influence my interpretation of the painting.	I can use the information provided to influence my response, and then use it to refer to previous knowledge.

Table 4: Distribution of responses for the self-reported understanding of the artworks divided by condition. The scale rating describes users' reported understanding with 1 being low and 7 being high understanding of the artwork. A chart version can be seen in Appendix B.

Scale	Social Media	Digital Collection	AI
1	83.33%	0%	16.67%
2	61.54%	30.77%	7.69%
3	30%	40%	30%
4	27.78%	27.78%	44.44%
5	28.57%	23.81%	47.62%
6	24%	44%	32%
7	0%	66.67%	33.33%

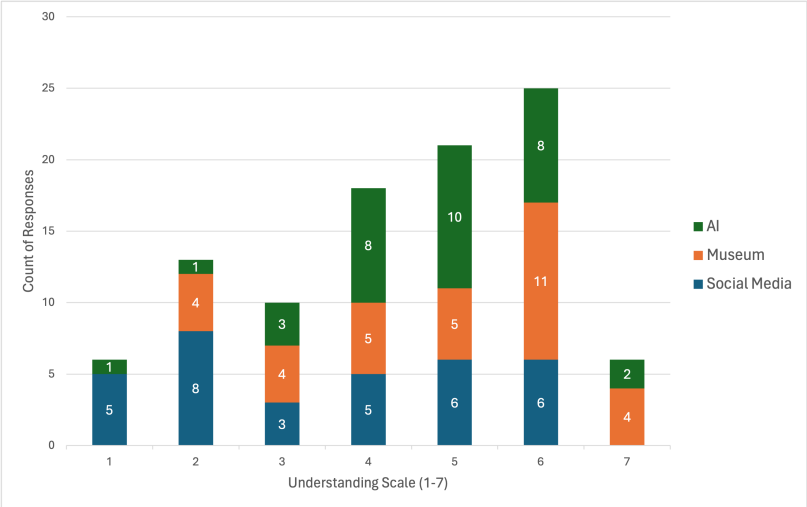


Figure 4: A larger chart representing the counts of responses to the Likert question asking user how much they understood the artwork.

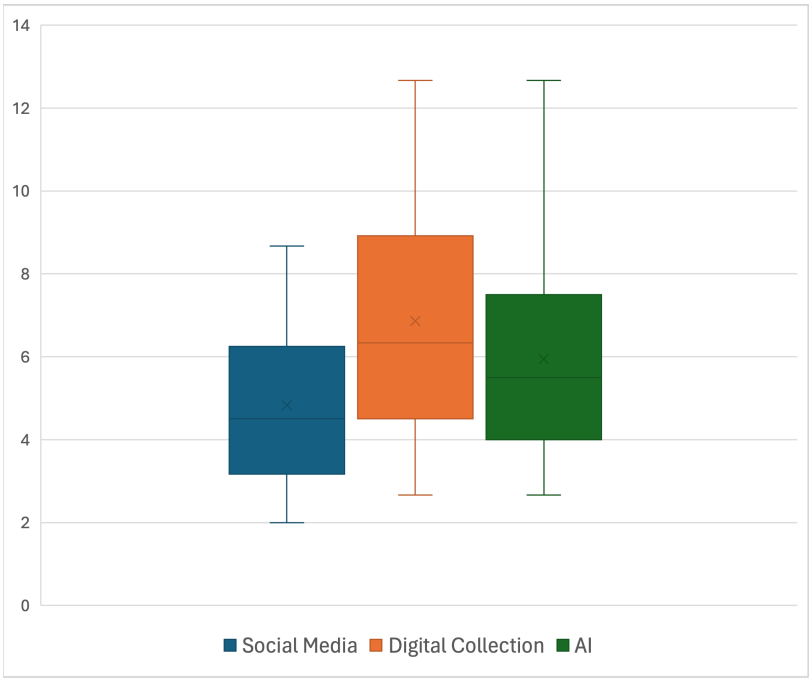


Figure 5: This boxplot displays the quartiles of difficult words used by the participants across each condition. While these are significantly different individual difference between conditions could not be specified by the Tukey post-hoc test.

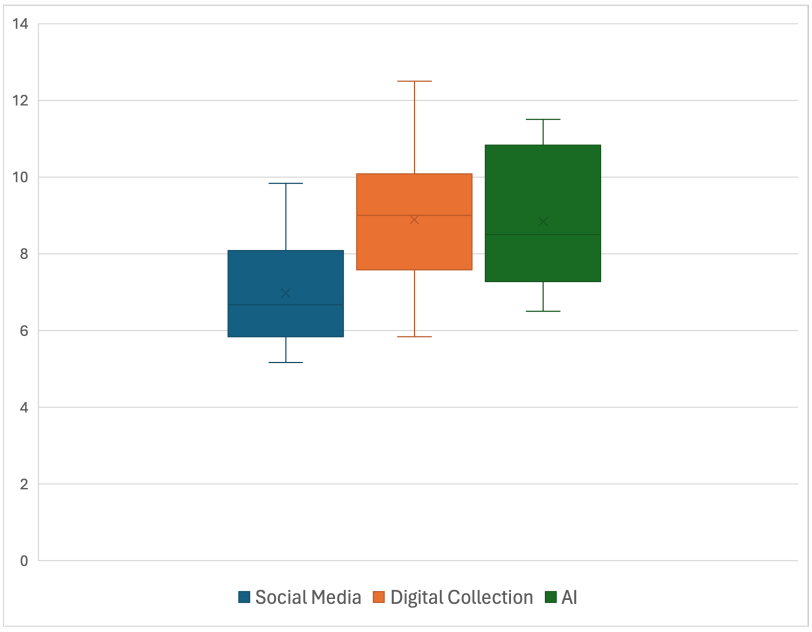


Figure 6: This boxplot displays the quartiles of grade level of the writing made by participants across each condition. These measures are different with both the Digital Collection and AI condition having a higher grade level than the social media condition.

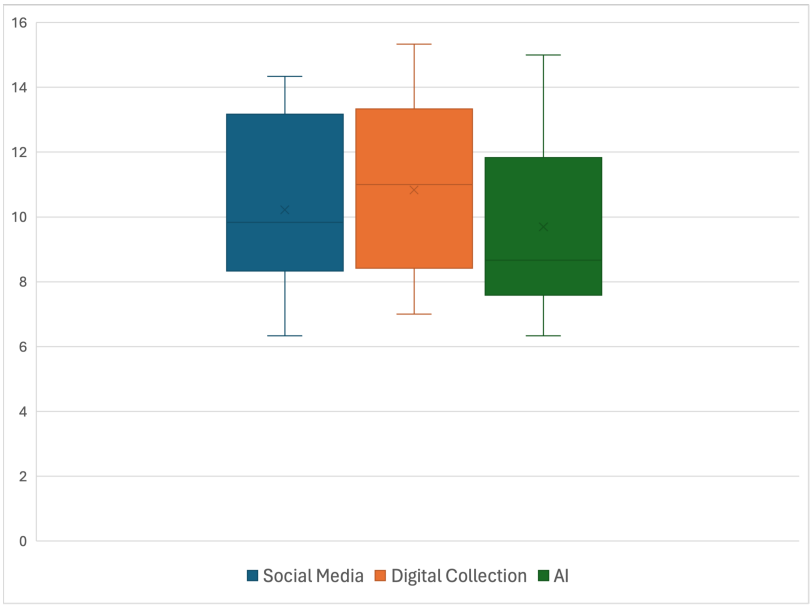


Figure 7: This boxplot displays the quartiles of the rubric score of the participants across each condition. These measures were not found to be significantly different across conditions.