ArtiQ: An Artist-in-the-Loop Chatbot for Question Answering in Art Galleries

Jinyu Liu
Durham University
Durham, United Kingdom
vvmj34@durham.ac.uk

Effie L-C Law
Department of Computer Science
Durham University
Durham, United Kingdom
lai-chong.law@durham.ac.uk

Hubert P. H. Shum
Durham University
Durham, United Kingdom
hubert.shum@durham.ac.uk

Abstract

The growing use of chatbots in art galleries raises design challenges in responding to the diverse questions that viewers pose about artworks. This study investigates how artist-in-the-loop workflows can be designed to address these challenges by structuring collaboration between chatbots and artists in responding to viewers' questions about artworks. We implemented this approach in ArtiQ, an art chatbot designed to operationalise the proposed artist-in-the-loop workflows by routing questions and enabling artists to refine chatbot responses, thereby allowing empirical examination in real-life gallery settings. We evaluated ArtiQ through a multi-site field deployment across four art galleries, involving 4 artists and 51 viewers. Adopting a mixed-methods approach, we integrated interaction logs of 420 viewer questions, viewer questionnaires (per-answer helpfulness ratings and the BUS-11 usability scale), and semi-structured interviews with artists and a subset of viewers. Quantitative analyses examined routing performance, viewers' ratings of chatbot responses across the two pathways, and the effort involved in artist editing, while qualitative analyses characterised how viewers and artists experienced the workflow. Results from the deployment demonstrate that artist-in-the-loop workflows can help maintain the quality of chatbot responses for viewer questions by retaining human interpretive control through selective artist involvement when interpretive expertise is warranted. These findings offer design implications for artist-in-the-loop systems, informing how LLM-powered chatbots can be integrated into cultural settings while preserving human authorship and interpretive integrity.

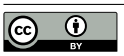
CCS Concepts

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **Empirical studies in HCI**;

Keywords

Art chatbots, Art appreciation, Artist-in-the-loop systems, Human-AI collaboration, Human-centered AI

ACM Reference Format:
Jinyu Liu, Effie L-C Law, and Hubert P. H. Shum. 2026. ArtiQ: An Artist-in-the-Loop Chatbot for Question Answering in Art Galleries. In *Proceedings of the 14th Nordic Conference on Human-Computer Interaction (NordiCHI '26), October 03–07, 2026, Vaasa, Finland*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3829807.3829831



NordiCHI '26, Vaasa, Finland

ACM ISBN 979-8-4007-2373-5/2026/10
https://doi.org/10.1145/3829807.3829831

1 Introduction

Art appreciation is widely understood as an active and dialogic process in which viewers construct meaning through questioning and interpretation, often accompanied by emotional responses to artworks [12]. In gallery settings, viewers often seek not only factual information but also insights into artistic intention, symbolism, and creative process [43]. Such questioning opportunities play a central role in shaping how viewers engage with and reflect on art [22]; however, sustained dialogue with artists is often limited in practice. As a result, many viewers leave exhibitions with unanswered or only partially explored questions [14]. In response to this gap, art institutions have increasingly adopted chatbots to support viewer engagement by providing scalable, on-demand access to information about artworks and exhibitions [8, 54]. While chatbots are effective at delivering descriptive or factual content, prior research suggests that interpretive and expressive questions remain difficult to address through fully automated systems [36, 51]. These limitations are particularly salient in art appreciation contexts, where misinterpretation may undermine artistic intent or interpretive integrity [12, 14, 19]. This tension highlights a key design challenge: how to leverage chatbot assistance in galleries without displacing the human expertise essential to artistic interpretation.

Prior work on conversational agents (chatbots[1]) and hybrid human–AI workflows in domains such as healthcare has proposed mechanisms such as question routing and draft-then-refine collaboration to balance automation with human expertise [2, 53]. Related museum-oriented systems have also explored human-in-the-loop approaches in adjacent contexts, such as museum touring and artifact description generation [27, 46]. However, these studies do not examine how such mechanisms function in art appreciation settings centred on viewer-authored questions, artist-led refinement of chatbot responses, and the preservation of interpretive authority. Moreover, prior work in economics [21] and other domains, as summarised in a recent scoping review [56], has predominantly examined such workflows in laboratory settings or from the perspective of a single stakeholder, leaving open questions about how they function in real-world, multi-stakeholder settings. In art contexts, there is limited empirical evidence on how these workflows function in situ, how viewers experience different response pathways, and how

[1]Although some researchers distinguish between conversational agents and chatbots based on factors such as capability, autonomy, or embodiment (e.g., [16, 34]), the terms are often used interchangeably in both research and practice to denote systems that support natural-language interaction. Following prior work (e.g., [15, 35]), we adopt this interchangeable use for clarity.

Figure 1: Example artworks used in the field deployment, illustrating three common thematic subjects in traditional Chinese ink painting: (left) flowers-and-birds, (middle) landscape, and (right) figurative composition. All artworks shown were part of the exhibitions where ArtiQ was deployed, and permission was granted by the artists for their inclusion in this paper.

artists engage with and benefit from chatbot-generated drafts as part of their interpretive work. As a result, it remains unclear how artist-in-the-loop workflows (cf. human-in-the-loop [38, 52]) perform in real-life gallery environments for art appreciation, where viewer questions vary widely in content and interpretive depth. Addressing this gap requires examining not only chatbot behaviour but also the experiences and efforts of both viewers and artists within an integrated interaction workflow. This paper investigates artist-in-the-loop workflows for handling diverse viewer questions in gallery settings through the following research questions (RQs). The RQs are ordered to reflect the logic of the workflow and study procedure.

- RQ1: How effectively does the system's routing mechanism support decisions about when viewer questions should be handled by the chatbot or escalated to the artist?
- RQ2: In real-life gallery settings, how do viewers evaluate the immediate chatbot responses they receive across the system's two routing pathways?
- RQ3: Under what conditions is human editing justified within the chatbot-artist hybrid workflow, and what factors shape the effort required for artists to refine chatbot-generated drafts?
- RQ4: How do viewers perceive the quality of artist-edited chatbot responses presented after their gallery visit?

To address these research questions, we conducted a field study centred on ArtiQ, an artist-in-the-loop art chatbot designed to support diverse viewer questions in gallery settings. The system draws on a previously proposed design framework for such interactions [33]. Rather than positioning ArtiQ as a standalone technical contribution, it serves as a deployable instantiation that enables us to examine key aspects of hybrid human–AI workflows, including routing, draft-based collaboration, and the practical implications of viewer-facing deployment in gallery settings. The ultimate long-term goal of our research work is to sustain public interest in artistic cultural heritage, which invites diverse interpretation and appreciation, by fostering meaningful and engaging dialogues between artists and viewers through a Q&A chatbot.

To evaluate the Q&A chatbot ArtiQ, we conducted a study in the domain of traditional Chinese ink painting, an artistic tradition characterised by strong emphasis on symbolism, expressive brushwork, and artist intention [7]. Rather than prioritising surface-level visual details, Chinese ink painting often emphasises the expression of inner states, temperament, and philosophical ideas [6], leading viewers to engage with artworks through a wide range of questions, from basic technique to interpretive and emotional inquiry. These characteristics make traditional Chinese ink painting a particularly suitable domain for examining how diverse viewer questions emerge and how chatbots and human artists should collaborate. Across the four participating galleries involved in the study (Section 3), the exhibited works spanned several well-established thematic subjects in traditional Chinese painting, such as landscapes, flowers-and-birds, and figurative compositions (see Figure 1).

Across the four sites, immediate chatbot responses and later artist-refined responses played complementary roles within the layered workflow. Automated replies supported viewers' timely access to initial responses to their questions, while follow-up interviews with the artists and a subset of participants suggested that artist involvement could add interpretive depth, emotional resonance, and contextual grounding to chatbot-generated drafts. Results indicated (Section 4) that the routing mechanism was generally effective in directing questions toward appropriate forms of response; however, it also revealed nuanced boundary cases where expectations for artist involvement varied. Importantly, editing drafts created by the chatbot required substantially less effort than composing responses from scratch, with the efficiency gains depending on how well the drafts aligned with artists' intentions,

highlighting the practical value and limitations of draft-based collaboration in interpretive domains. These findings suggest that artist-in-the-loop workflows offer a practical way to selectively offload some question-answering work to LLM-powered chatbots in art galleries without displacing human authorship (Section 5). By selectively structuring when and how artists are involved through layered responses, routing decisions, and draft-based refinement, such systems can balance scalability with interpretive integrity, supporting meaningful viewer engagement while making artist participation sustainable over time. This paper makes the following contributions:

- Insights from a multi-site field study of an artist-in-the-loop art chatbot (ArtiQ) in real-life gallery settings, exploring viewer-chatbot interactions, question routing, and viewers' perceptions of immediate chatbot responses during on-site use.
- An in-depth understanding of the asynchronous artist-refinement stage, examining how artists engaged with chatbot-generated drafts and how selected viewers subsequently perceived the artist-edited responses within the layered workflow.
- Empirical evidence on draft-based human-AI collaboration in an interpretive art context, showing how the quality and alignment of chatbot-generated drafts relate to artists' editing effort and the efficiency of collaborative response production.
- Design insights for artist-in-the-loop chatbot systems, translating empirical findings from ArtiQ into guidance on how layered responses, selective routing, and draft-based collaboration can support the integration of LLM-powered chatbots into art appreciation settings while preserving artistic authorship and interpretive integrity.

2 Background and related work

2.1 Art Appreciation, Viewer Engagement, and Questioning Behaviors

Art appreciation is widely understood as an active meaning-making process rather than a passive reception of visual stimuli [43]. Viewers interpret artworks through a combination of prior experience, cultural knowledge, and emotional response. The same piece can, therefore, be understood in different ways by different people [13]. Contextual information can further shape aesthetic engagement, influencing what viewers notice and how they construct narratives around an artwork [10]. Emotional or affective responses have also been discussed in prior work as part of viewers' broader aesthetic experience, contributing to how artworks are encountered and reflected upon [43].

Within this appreciation process, viewer engagement is closely tied to opportunities for dialogue and questioning. Prior research suggests that dialogic approaches can deepen viewer engagement by moving beyond one-directional explanations from artists or institutions and instead foregrounding viewers' opportunities to articulate and pursue their own questions [12]. Interpretive questions that invite viewers to reflect on emotional responses or artistic intention can encourage viewers to revisit the work, compare perspectives, and situate it within their own experiences [20]. More broadly, viewers have increasingly been described as moving from passive recipients to active participants who expect to explore, comment, and co-construct meaning with institutions and artists [32].

Prior work in gallery and museum contexts suggests that viewers' questions are heterogeneous, ranging from requests for basic factual information to more interpretive, experiential, and process-oriented inquiries (e.g., [33]). Such variation reflects different ways of engaging with artworks, including grounding understanding in core information, probing symbolic or emotional meaning, and seeking guidance on how a work was created or should be approached [12].

2.2 Conversational Agents in Art Settings and the Limits of Automated Interpretation

Art institutions and cultural spaces have increasingly adopted chatbots to support access, engagement, and learning. These applications are commonly introduced to provide scalable viewer guidance, deliver contextual information on demand, and enable more personalised exploration of exhibitions [8, 54]. By offering real-time responses to viewer queries, chatbots can provide traditional interpretive materials and create more interactive art-viewing experiences [8]. Such applications reflect broader efforts to enhance inclusivity and participation through digital tools within art contexts [18]. In prior work [4], user experience with conversational agents has been assessed with validated usability instruments such as the Chatbot Usability Scale (BUS-11), an 11-item questionnaire that captures users' perceived clarity, usefulness, engagement, and overall interaction quality with human-AI chatbots. The BUS-11 shows a stable factor-structure, high internal consistency (Cronbach's alpha >0.9), and convergent validity with other usability measures such as the UMUX-Lite [29], making it a robust tool for assessing user satisfaction and perceived usability of chatbot systems [4, 5].

Technically, chatbots deployed in art settings draw on a spectrum of conversational agent approaches, ranging from earlier rule-based systems, which rely on predefined scripts and predictable responses, to more recent generative systems based on LLMs that produce open-ended responses [9, 49]. While LLM-driven chatbots can provide more flexible and conversational interaction, they also introduce risks, including hallucinations, inaccuracies in contextual information, and biases inherited from training data [28, 31]. These risks are particularly consequential in cultural and artistic contexts, where misinterpretation can distort intent or propagate misleading narratives [51]. As a result, prior research in art emphasises the importance of human expertise, especially when interpretive nuance or cultural sensitivity is required [12, 14].

Trust and perceived suitability further differentiate human experts and chatbots in art interpretation. Prior work in cultural heritage suggests that interpretive authority in cultural contexts is closely tied to human expertise [8, 51], particularly when responses involve emotional nuance, symbolism, or contextual judgment. While chatbots can reliably deliver factual information, their limited capacity for emotional understanding and absence of situated cultural experience may make them less suitable for subjective or expressive interpretation [3, 37, 42]. These tensions motivate

approaches that combine automated efficiency with human interpretive depth.

The boundaries of automated interpretation are thus shaped by the nature of the content being addressed. Objective information, such as basic factual details, is well aligned with automated delivery and benefits from the speed and consistency of AI systems [19, 38]. In contrast, interpretive questions require contextual reasoning, emotional attunement, and sensitivity to artistic intention, which current LLMs struggle to replicate reliably [41]. These limitations highlight the importance of hybrid approaches that leverage AI for factual clarity while engaging human experts in meaning-making tasks, particularly in domains that are open to interpretation, such as art appreciation.

2.3 Hybrid Human-AI Workflows for Interpretation: Routing, Drafting, and Collaborative Answering

Hybrid human-AI workflows have gained increasing attention across knowledge-intensive domains involving interpretation and meaning-making, where automated systems and human experts contribute complementary strengths [25, 39, 44]. Prior work on human-AI collaboration highlights how AI systems can support tasks such as information delivery and content generation, while humans retain responsibility for contextual judgement, creativity, and domain expertise [1, 21, 56]. A common pattern in such workflows is for AI models to produce an initial draft that human experts subsequently refine; an approach shown to improve efficiency while preserving interpretive control [2]. When AI contributions are framed as provisional starting points rather than final outputs, human collaborators remain in control of the interpretation and context of the response, while benefiting from the speed and consistency of automated assistance.

Museum chatbot research spans rule-based systems (MUSU [24, 45]), knowledge-graph approaches (MuBot [47, 48]), and recent generative agents [50]. Curator or artist involvement remains largely limited to pre-generation stages, despite concerns about reliability and curatorial authority. Post-generation oversight of AI outputs is underexplored. Related hybrid work, including expert refinement of AI-generated descriptions [27] and human-in-the-loop touring with quality of experience (QoE)-based routing [46], does not address viewer-authored questions in art galleries or selective routing between chatbot-only and artist-refined responses that preserves interpretive authority.

Within such hybrid workflows, and especially in draft-then-refine arrangements, the usefulness of AI-generated drafts is closely tied to their quality. High-quality drafts can reduce cognitive load by minimizing low-level editing, enabling human contributors to focus attention on deeper conceptual or expressive refinements [26]. In contrast, drafts that are overly generic, inaccurate, or stylistically misaligned may increase revision effort, particularly when human experts must correct subtle interpretive errors or re-establish contextual grounding [11]. These findings highlight that draft generation is not merely a time-saving feature; rather, it affects the level of human labour and the extent to which collaborative systems can support expert-driven workflows.

Effective collaboration also depends on routing mechanisms that determine when AI is sufficient and when human involvement is needed [55]. Research on query classification and intent detection shows that systems benefit from identifying question type, ambiguity, or multi-intent structure in order to direct queries to appropriate response pathways [23, 53]. Such routing improves response quality, minimises user frustration, and ensures that subjective or context-dependent questions are escalated to human experts rather than handled solely by automated models [1, 53]. These approaches are particularly important when misclassification risks undermine trust or interpretive accuracy [40].

3 METHODS

3.1 System Design and Technical Setup

ArtiQ is an artist-in-the-loop chatbot developed for use in gallery settings, designed to support a variety of viewer questions. The system differentiates between questions focused on factual aspects of artworks and those that are more interpretive or expressive, directing them to either a **chatbot-only (CO)** or a **chatbot–artist hybrid (CAY)** response pathway. For clarity throughout the paper, we refer to factual questions as **Artwork Factual (AF)** and interpretive or expressive questions as **Interpretive and Expressive (IE).** This distinction aligns with prior work on categorising viewer questions in gallery contexts [26]. The technical components described below function as infrastructural elements that support the study.

3.1.1 Routing Mechanism. At the core of ArtiQ is a routing mechanism (Figure 2) that distinguishes questions grounded in descriptive or technique-oriented aspects of the artwork (i.e., AF) from those involving interpretation, intention, or meaning (i.e., IE). When a viewer submits a question, the chatbot system applies a lightweight on-device classifier trained on a previously established taxonomy of viewer questions to categorise the question as AF or IE. Questions classified as AF are answered directly under the CO pathway whereas questions classified as IE are routed to the CAY pathway, where the chatbot provides an initial draft response to the viewer first and the artist later refines the response. This routing mechanism ensures that AF questions are handled efficiently by the chatbot system while reserving interpretive and expressive questions for the artist.

To support reliable routing under real-world deployment constraints, ArtiQ adopts a dual-layer classification strategy. Upon question submission, the system first applies a lightweight Naïve Bayes classifier trained on the same question taxonomy. The classifier outputs both the predicted label (AF vs. IE) and a confidence score. Predictions with confidence > 0.7 are accepted as the final routing decision, a threshold chosen empirically in pilot testing to balance routing accuracy and system responsiveness. For low-confidence cases (≤ 0.7), ArtiQ escalates the question to a secondary classifier based on the Qwen LLM (provided by Alibaba Cloud), which performs a second-pass AF/IE categorisation. This LLM-based classifier is adapted using task-specific instructions, including concise definitions of AF and IE and representative examples, and its output is used as the final label. We used the LLM-based fallback only for low-confidence cases to reduce misrouting while keeping

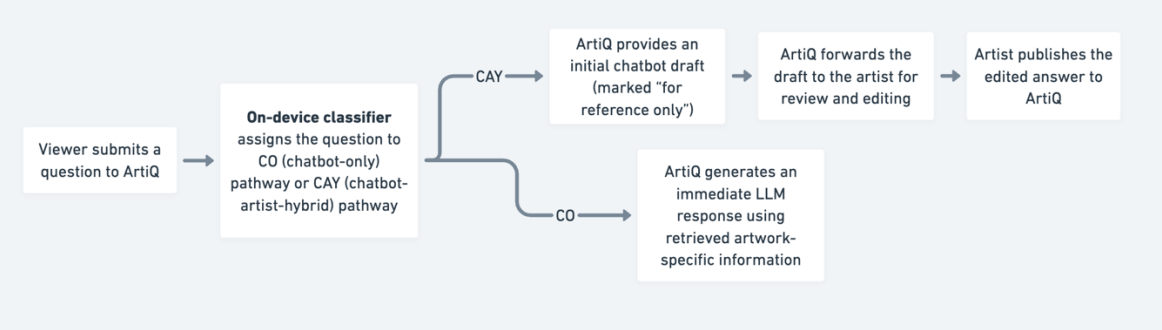


Figure 2: ArtiQ's routing mechanism and response generation process. All responses appear in the viewer's 'My Questions' panel. The chatbot logs all interactions for later quantitative and qualitative analysis.

on-device latency low. All viewer questions are entered in Chinese; no explicit length limit is imposed, allowing and encouraging viewers to express their queries about the artworks freely.

3.1.2 Response Generation. ArtiQ supports two response pathways corresponding to the routing outcomes (Figure 2). In the CO pathway, AF questions are answered immediately by the chatbot using retrieval-augmented generation (RAG) grounded in curated contextual information [30]. This provides consistent and concise explanations while giving viewers quick access to artwork details. In the CAY pathway, IE questions receive a chatbot-generated initial draft response that is explicitly marked "for reference only." ArtiQ notifies the viewer that the question has been forwarded to the artist for further refinement. Through the artist's interface, artists review and refine the draft to ensure that the final answer reflects their creative intent and expressive choices before publishing it to the viewer.

Responses in both pathways are generated using the Qwen-Turbo LLM via an API-based deployment. Prior to deployment, participating artists provided structured background materials for each artwork, which were used to initialise a curated knowledge base (Section 3.5.2). These materials, covering artwork information, artist statements, technique glossaries, and exhibition information, were embedded and stored in a vector store. When a viewer submits a question under the CO pathway, the question is embedded and used to retrieve relevant documents from the vector store, scoped to the corresponding artwork. The retrieved materials are then injected into the prompt as contextual evidence for RAG, encouraging factual, artwork-specific responses and reducing unsupported elaboration. This process grounds the model's responses in artwork-specific documents and is part of the system's core functionality.

3.1.3 Interfaces. ArtiQ comprises three role-specific interfaces that support the proposed workflow across all deployment galleries: viewer, artist, and admin interfaces. Figure 3 illustrates the viewer and artist interfaces used during the interaction workflow (Note: a video demonstrating the end-to-end interaction flow is provided in the accompanying supplementary material).

- Viewer interface: Viewers, after encountering the artwork in its physical form on display, can refer to its digital version when formulating questions and submit them to ArtiQ. They then read responses through the viewer interface (Figure 3). All answered questions are displayed in the My Questions panel, where CAY cases show both the chatbot's initial draft and the artist-edited final answer.
- Artist interface: Artists receive routed questions, review the chatbot's initial draft, and refine the content before publishing the final response. Artists are also asked to indicate whether forwarding this question is necessary, providing a ground-truth check of the chatbot's classification performance. The interface also includes a five-star rating widget for artists to evaluate the chatbot's initial draft response, which supports later analysis of human-AI collaboration.
- Admin interface: Researchers use the admin interface to manage viewer registration, monitor routing behaviour during deployment, and export interaction logs for analysis.

The chatbot system automatically logged all interaction events, including question texts, routing outcomes, response drafts, final answers and timestamps.

3.2 Exhibition Sites and Artwork Selection

We conducted a multi-site field deployment of ArtiQ across four gallery spaces in China located in four different cities that were geographically distributed. The galleries were similar in scale, ranging from approximately 50 to 100 m^2, and attracted a modest but steady stream of viewers (roughly 20-50 per day). These small to mid-sized exhibition settings provided naturalistic yet broadly comparable environments for deploying the chatbot, enabling us to examine how ArtiQ operates within typical gallery viewing conditions.

Each gallery featured one participating artist (A1, A2, A3 and A4), all of whom are painters specialising in traditional Chinese painting. The galleries involved in the deployment primarily exhibit this art form, ensuring that both the viewing context and the interaction content were centred on a coherent artistic domain. For the study, we collaborated with each artist to select four to five of their artworks currently on display. All artworks were included with the artists' explicit permission, and their use within ArtiQ was authorised solely for research and exhibition purposes. The artwork selections followed three criteria:

- the artworks were representative of the artist's style and technique within traditional Chinese painting;

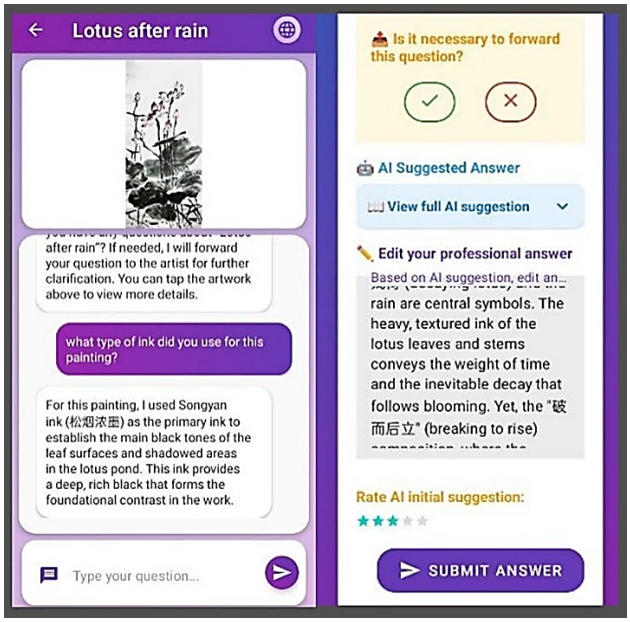


Figure 3: Viewer and artist interfaces used during the interaction flow (the admin interface is not shown). (Left) Viewer interface receiving chatbot-generated responses. (Right) Artist interface displaying the chatbot's initial draft responses and tools for editing and rating. For readability, the interface screenshots shown here and the video in the supplementary material were translated into English and are illustrative demo views rather than exact reproductions of the original Chinese-language tablet interface used in the gallery deployment.

- the artworks offered sufficient visual and thematic richness to elicit both descriptive and interpretive questions from viewers;
- the artworks were accompanied by sufficient background materials to allow the art chatbot system to generate contextually grounded responses.

Across all four sites, ArtiQ was deployed on a tablet positioned within the exhibition area, allowing viewers to browse the artworks and submit questions directly in front of the pieces or elsewhere in the gallery, as the digital versions of the artworks are available in the chatbot system. This setup embedded the interaction within viewers' natural appreciation flow and avoided the artificiality of a laboratory setting, reinforcing the study's emphasis on the feasibility of artist-AI collaboration in real-life gallery environments.

3.3 Participants

Two groups of participants were involved in the study: gallery viewers (N = 51 in total across the four gallery sites), who interacted with ArtiQ during their gallery visit, and the four artists (A1-A4) whose artworks were displayed at the gallery sites. All participation was voluntary without compensation or reward. Ethical approval was obtained from the host institution, and informed consent was collected from all participants.

Participation did not require any prior knowledge of traditional Chinese painting or chatbot technologies. Viewers were free to explore the exhibition and use the chatbot as part of their visit. To characterise the viewer sample, a demographic questionnaire collected viewers' age, gender, educational background, familiarity with traditional Chinese painting, and gallery-visiting frequency. After completing the questionnaire, viewers could voluntarily leave contact information if they were willing to be contacted for a possible follow-up interview. All of them left their email addresses. The 51 viewers ranged in age from 19 to 64 years (M = 34.8, SD = 10.7), and included 31 women and 20 men. Additional descriptive statistics are shown in Table 1. Because recruitment took place in real-life gallery settings, the resulting sample reflected the characteristics of the viewers who happened to be present.

The four participating artists are established practitioners of traditional Chinese painting, each with an active exhibition at their respective gallery sites. All have substantial professional experience in creating and exhibiting traditional Chinese paintings. These artists were invited because of their sustained practice within this domain and their familiarity with communicating artistic intent to the public. Their role in the study is described in detail in Section 3.5.

3.4 Instruments

To capture both the interaction dynamics and participants' experiences, we collected the following quantitative and qualitative measures (see Table 2). All questionnaires except BUS-11 were developed in-house because no existing instruments adequately addressed the specific constructs we needed to measure. For routed questions, additional variables including artist ratings, viewer ratings, response times, and question characteristics were recorded to examine the performance of the chatbot and human-AI collaboration.

3.5 Procedure

Across the four evaluation setups, the study procedure consisted of three stages (see Figure 4). These stages included the on-site viewer study, artist off-site editing workflow, and follow-up interviews with the artist and a subset of viewers.

3.5.1 Viewer Procedure (On-site Study). Upon entering the gallery, viewers were briefly asked by gallery staff whether they were interested in taking part in the study. Those who agreed provided informed consent and first completed a demographic questionnaire. Then, viewers checked a printed instruction sheet outlining how to use the tablet-based chatbot and the steps involved in the study. Gallery staff did not provide any assistance during the experiment, and the exhibiting artists were not present to avoid any influence on the viewers' behaviour. Viewers browsed the artworks freely and submitted questions directly through the tablet. Because each gallery consisted of a compact, open viewing space in which all exhibited works were visible simultaneously, the spatial arrangement of artworks did not impose a fixed interaction order. After receiving each response, viewers provided an immediate rating through the interface. Once they finished the Q&A-based interaction, viewers were asked to complete a questionnaire that included the BUS-11

Table 1: Viewer educational background, familiarity with traditional Chinese painting, and gallery-visiting frequency (N = 51)

Variable	Category	n	%
Educational level	Below undergraduate	13	25.6
	Undergraduate	19	37.2
	Postgraduate or above	19	37.2
Familiarity with traditional Chinese painting	Not familiar	10	19.6
	Slightly familiar	16	31.4
	Moderately familiar	15	29.4
	Very familiar	10	19.6
Gallery-visiting frequency	Once a year or less	9	17.6
	About once every six months	11	21.6
	About once every three months	20	39.2
	Once a month or more	11	21.6

Table 2: Instruments and variables used in the study

Instrument / Variable	Respondent	Description	Type / Scale
Chatbot Usability Scale (BUS-11)	Viewer	A validated 11-item questionnaire completed at the end of the on-site interaction to assess overall satisfaction with the chatbot interaction they experienced, covering clarity, usefulness, engagement, and overall experience (see Section 2.2). It was not used to evaluate the artist-refined stage of the workflow.	Continuous
Per-answer helpfulness rating (viewer_rating)	Viewer	After each chatbot-generated response, viewers rated how helpful they found the response on a 5-point Likert scale.	Ordinal (1-5)
Open-ended reflections	Viewer	A post-study questionnaire consisting of four open-ended prompts asking what viewers liked most and least about the interaction, their expectations regarding when artist input would be appropriate, and suggestions for improvement.	Qualitative
Chatbot initial draft rating (artist_rating)	Artist	For each routed question, artists provided a 5-point Likert-scale rating, reflecting their overall assessment of the quality of the chatbot-generated draft response.	Ordinal (1-5)
Escalation judgement (should_escalate)	Artist	Artists indicated whether each question should have been escalated to them (Yes/No), providing ground-truth labels for evaluating the routing mechanism.	Binary
Edit time (edit_time)	Artist	Time taken to revise the chatbot-generated initial draft.	Continuous (seconds)
Write time (write_time)	Artist	Time taken to write an answer from scratch.	Continuous (seconds)
Question length (question_text_length)	System	Length of the viewer's question text.	Continuous
Routing label (predicted_label)	System	System-predicted routing outcome (CO / CAY).	Categorical
Question category (question_category)	System	Question category assigned to the viewer question (AF / IE)	Categorical

satisfaction measure and four open-ended questions about their experience (Table 2).

3.5.2 Artist Preparation and Workflow. Before the chatbot deployment at each gallery, each artist provided textual background materials about the selected artworks (e.g., artist descriptions and exhibition notes) to initialise the chatbot's knowledge base. The materials were collected through a structured form to ensure that the art chatbot system had consistent contextual information for generating responses. After all viewer interactions had concluded at each site, artists completed their part of the workflow through the dedicated artist interface. For every question routed to the CAY pathway, the interface displayed the viewer's original query alongside the chatbot-generated draft response.

To examine the effort involved in different forms of human intervention, artists completed two related tasks for every routed question in a fixed order. First, they wrote a fresh answer to the viewer's question from scratch; this response was not shown to

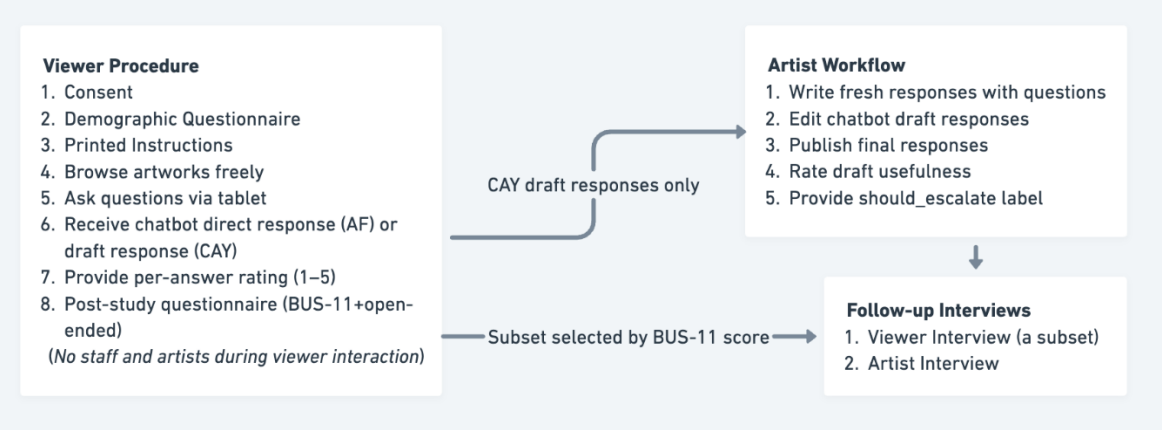


Figure 4: Procedure of the multi-site evaluation, detailing viewer and artist workflows.

viewers and served solely as a baseline for comparing writing and editing time. At a later stage, they reviewed the chatbot-generated draft and edited it as needed to produce the final answer, which was published back to the viewer's channel.

This ordering was intentional. It ensured that both writing and editing times were collected for the same set of routed questions. Reversing the order was considered but not adopted because exposure to chatbot-generated drafts could prime how artists formulate their own responses, compromising the independence of the writing task. Writing first therefore provides a consistent baseline against which chatbot-assisted editing can be interpreted. To reduce immediate dependency, the two tasks were separated by approximately one day rather than performed consecutively. This interval helped limit short-term influence while accommodating the system's asynchronous, multi-site use. We acknowledge that this design does not capture editing behaviour in a fully naturalistic sense, as artists were not editing chatbot drafts from a blank starting point. Accordingly, the recorded times are interpreted as indicative of relative effort within this structured workflow, supporting a comparative and exploratory understanding of draft-based interaction rather than a precise measure of editing effort. Further work would be needed to examine editing practices in more ecologically unconstrained settings.

After completing each final response, artists rated the chatbot-generated draft and indicated whether the question should have been escalated to them. As artists completed the editing task asynchronously (i.e., not in real time as viewers submitted questions), edited answers were published only after viewers had left the gallery. As a result, viewers did not have the opportunity to read the edited answers on site (the tablet running ArtiQ remained in the gallery), unless they later took part in the semi-structured interviews described below.

3.5.3 Follow-up Interviews. In addition to the on-site study, follow-up interviews were conducted to further contextualise the viewer and artist experiences. After the artist workflow, twelve viewers (three from each gallery, sampled to include high, medium, and low BUS-11 scores) were invited to participate in a follow-up online interview. These interviews took place after the gallery visit. In each interview, participants were shown the artist-edited final answers corresponding to the questions they had submitted during the study, together with the original chatbot draft responses; these materials were retrieved by the researcher in advance and used as prompts for discussion. Participants were asked to reflect and comment on their perceptions of the workflow and the differences between each chatbot-generated draft and its artist-edited response.

Semi-structured interviews were also conducted with the four participating artists following the completion of all editing tasks. These interviews focused on their experience of the current workflow, including perceptions of draft quality, the appropriateness of routing decisions and factors influencing editing effort. All interviews were audio-recorded and transcribed for subsequent qualitative analysis.

The qualitative materials were analysed through an iterative coding process to identify recurring patterns relevant to the research questions, including perceived strengths and limitations of chatbot responses, expectations around artist involvement, and factors shaping editing effort. Initial coding was conducted by the first author and subsequently refined through discussion with another researcher with art expertise. These codes were then grouped into higher-level themes that informed the qualitative findings reported in Section 4.

4 RESULTs

In this section, we present our results structured around the four RQs (Section 1), including analyses of quantitative variables (Table 2) and relevant qualitative data.

4.1 RQ1: Routing Effectiveness

4.1.1 Quantitative Results. To evaluate how well the routing mechanism aligned with artists' intervention judgements within the current workflow, each question was assigned two labels: the system-predicted response pathway (CO vs. CAY) and the artist's judgement on whether the question should be escalated to a human artist (should_escalate, 0 = No, 1 = Yes). In this analysis, artist judgement served as a practical operational reference for whether artist involvement was warranted, rather than as an absolute ground truth

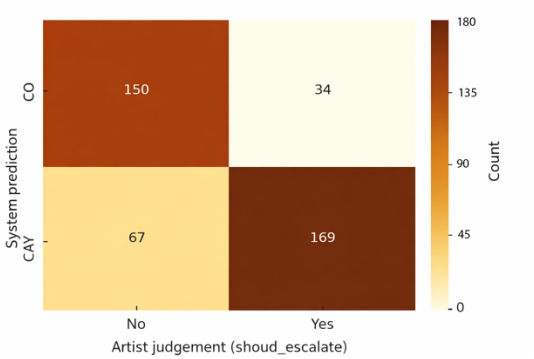


Figure 5: Confusion matrix comparing system routing predictions (CO vs. CAY) with artists' escalation judgements.

for viewer intent. By comparing these two labels, a 2×2 confusion matrix (Figure 5) was generated, where true positives (TP = 169) represent questions correctly escalated to artists, true negatives (TN = 150) are questions appropriately handled by the chatbot alone, false positives (FP = 67) are questions unnecessarily escalated, and false negatives (FN = 34) are those that should have been escalated but were not. From this matrix, key performance metrics were computed: accuracy = 0.760, precision = 0.716, recall = 0.833, and F1-score = 0.770. These values indicate that the classifier was able to route questions with a strong balance between precision and recall, successfully identifying most of the queries that required artist involvement for IE questions. The slightly higher recall suggests a conservative bias favouring escalation, which is desirable in our context to avoid missing IE questions.

4.1.2 Qualitative Results. Drawing on open-ended survey responses from all viewers and follow-up interviews with artists and a subset of viewers, the qualitative findings show that routing behaviour shaped interaction in two complementary ways. Most viewers appreciated that the chatbot was generally able to "know what needs the artist's input," especially when questions touched on meaning, personal experience, or artistic intention. As one viewer explained: "*For the deeper questions, it was right to let the artist further explain it. Only the artist knew what he felt when painting it.*" (P29). Artists similarly acknowledged that many handovers were appropriate. A3 noted: "*When viewers asked about my mood or why I placed blank spaces in the composition, these were exactly the kinds of questions that should be handled by the artist, and the chatbot system consistently forwarded them to me.*"

However, both follow-up interviewees and artists pointed to a small set of questions that fell into an ambiguous grey zone between factual clarification and interpretive inquiry. Two interviewees, for example, asked colour-related questions—"*Why is the fruit in yellow?*" (P25) and "*Why is the flower in orange?*" (P14)—which the chatbot system handled by providing a draft response and then forwarding the query. Yet their reactions diverged. P25 felt the escalation was unnecessary because the factual clarification "*was exactly the level of detail I wanted.*"

In contrast, P14 welcomed the artist's involvement, noting that colour choices often carry emotional or symbolic resonance that "only the artist could explain." Artist interviews revealed a similar divergence. A1 expressed frustration with receiving questions he felt were straightforward: "Some simple questions were sent to me, like 'Is learning traditional Chinese landscape painting difficult?' This could be directly answered by the chatbot." A3, however, viewed such queries as interpretive: "When someone asks whether learning painting is difficult, they're not only asking about skill level—they're asking about the artistic mindset and how to begin. I want to answer that myself." These cases suggest that routing appropriateness cannot be determined solely in terms of technical classification accuracy: it also depends on how viewer expectations and artists' thresholds for intervention are negotiated in ambiguous cases.

4.2 RQ2: Viewer Evaluation of Chatbot Responses

4.2.1 Quantitative Results. Analysis of all 420 viewer questions revealed clear differences in how viewers evaluated the helpfulness of immediate chatbot responses produced between the two major question types (AF, n=184 vs. IE, n=236). Overall, responses to AF questions received higher viewer ratings than responses to IE questions, with mean scores of M = 4.30 (SD = 1.02) for AF questions and M = 3.40 (SD = 1.55) for IE questions (Figure 6). Shapiro-Wilk tests confirmed that both groups deviated from normality (AF: $p < .001$; IE: $p < .001$); therefore, non-parametric testing was applied. A Mann–Whitney U test revealed a significant difference between the two question types (U = 14,480, Z = −6.196, $p < .001$), with higher mean ranks for AF questions (Mean Rank = 249.80) than for IE questions (Mean Rank = 179.86). Visual inspection of Figure 6 suggests a ceiling effect for AF questions, with ratings clustered at the upper end of the scale, whereas ratings for IE questions span a wider range, indicating greater variability in viewer experience. These findings suggest that the chatbot may not handle IE questions as effectively as AF questions, which is consistent with the system's design choice to route IE questions to artists. However, there may also be an experimental artefact: viewers were told that the chatbot would forward their IE questions to an artist, which could have acted as a proxy confidence cue, implying that the chatbot's response would be of lower quality than the artist's. This information may have primed or biased viewers to rate the responses to IE questions more negatively than they otherwise would have. This possibility cannot be ruled out, although the present data do not allow us to assess its influence systematically. Although participants could have been asked to justify their satisfaction ratings for each chatbot response, doing so would likely have been cognitively taxing and could have led to response fatigue.

To further verify whether the overall trend was consistent across individual gallery sites, Shapiro-Wilk tests were first conducted for both AF and IE question groups within each gallery: Gallery 1 (AF, n = 35; IE, n = 60), Gallery 2 (AF, n = 47; IE, n = 60), Gallery 3 (AF, n = 49; IE, n = 45), and Gallery 4 (AF, n = 53; IE, n = 71). As all tests revealed deviations from normality ($p < .001$), Mann-Whitney U tests were applied separately for each gallery. The results revealed a consistent directional pattern across all four sites, with responses to AF questions generally rated higher than responses to IE questions. Specifically, the difference was not statistically significant

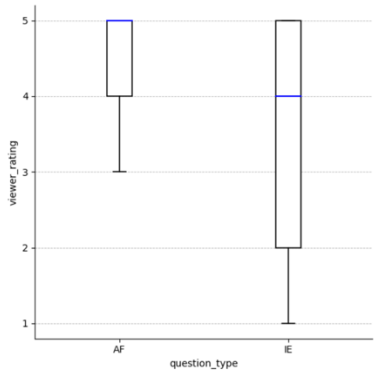


Figure 6: Distribution of viewer ratings across AF and IE questions.

in Gallery 1 ($Z = -1.38$, $p = .169$), but was highly significant in Gallery 2 ($Z = -3.80$, $p < .001$), Gallery 3 ($Z = -3.76$, $p < .001$), and Gallery 4 ($Z = -3.71$, $p < .001$). The non-significant result in Gallery 1 is consistent with its question profile, which included a higher proportion of questions positioned closer to the boundary between factual clarification and interpretation, thereby reducing the observed rating gap between AF and IE questions. These findings indicate that the observed difference between AF and IE questions was consistent across the four gallery sites included in this study.

We assessed the internal consistency of the BUS-11 scale in our dataset. The results indicated good internal reliability (Cronbach's $\alpha = .818$) (i.e., comparable to the value reported in [5]), supporting the use of the aggregated BUS-11 score as a measure of overall chatbot satisfaction with the on-site chatbot interaction. At the participant level, correlations between the mean viewer ratings and the overall BUS-11 satisfaction scores were examined to assess whether ratings of individual chatbot responses aligned with broader evaluation of the chatbot. Each participant's mean viewer rating was computed by averaging all of their per-answer helpfulness ratings provided across both AF and IE questions, as every chatbot response was evaluated using the same 5-point scale. The number of ratings contributing to each participant's mean ranged from 4 to 12 ($M = 8.24$, $SD = 1.69$), reflecting how many questions each viewer chose to ask. As reported below, the average number of questions per participant did not differ significantly across the four galleries, indicating that the number of ratings contributing to each participant's mean was broadly comparable across sites. Results of the Shapiro-Wilk test indicated that both variables followed approximately normal distributions ($p > .05$). A Pearson correlation analysis was therefore performed. The results revealed a significant positive correlation between participants' BUS-11 total scores and their mean viewer ratings ($r = .551$, $p < .001$, $N = 51$). This moderate correlation indicates that participants who gave higher per-answer ratings also reported higher overall satisfaction with the chatbot. In other words, viewers' assessments of individual chatbot response quality during question-answer interactions were moderately consistent with their overall perception of the on-site chatbot experience. The scatter plot (Figure 7) illustrates this relationship, showing a positive upward trend consistent with a moderate association between question-level satisfaction and overall user experience.

To confirm that these findings were not affected by uneven data distribution, additional analyses were conducted to examine question frequency across galleries and artworks. For the gallery-level analysis, we computed the average number of questions per participant based on the artworks they chose to engage with in each gallery, as viewers were free to ask questions about any subset of the artworks on display rather than being required to interact with all artworks. The four galleries showed highly comparable participant-level question counts: Gallery 1 ($M = 7.9$, $SD = 1.1$), Gallery 2 ($M = 8.6$, $SD = 3.4$), Gallery 3 ($M = 8.5$, $SD = 1.5$), and Gallery 4 ($M = 8.3$, $SD = 1.3$), indicating that viewers across sites asked similar numbers of questions. Shapiro-Wilk tests indicated approximately normal distributions across groups ($p > .05$ for Galleries 2 and 4; slight deviations for 1 and 3 were acceptable given sample sizes). A one-way ANOVA on the average number of questions per participant across the four galleries showed no significant differences ($F(3, 47) = 0.29$, $p = .835$, $\eta^2 = .018$). At the artwork level, across the 17 artworks included in the deployment, descriptive statistics indicated that each artwork received between 21 and 32 questions ($M = 24.71$, $SD = 3.70$), suggesting a relatively balanced question distribution across artworks.

4.2.2 Qualitative Results. Across all four exhibitions, the open-ended responses showed a consistent pattern that complements the quantitative findings: viewers described the chatbot as fast, accurate, and helpful for clarifying factual or technical aspects associated with AF questions. Many respondents characterised the chatbot's replies as providing a "stable baseline" for understanding techniques and materials. As one viewer wrote, "*It directly answered most of my technique questions very clearly, like how the ink layering works.*" (P12). Several others noted that the chatbot often directed their attention to visual details they had previously overlooked—for example, "*The chatbot told me to look at the S-shaped movement in the composition. I immediately wanted to ask more about why the artist arranged the elements this way.*" (P42). By contrast, the open-ended responses were less positive about IE initial draft responses. Some viewers described these responses as useful but incomplete, expressing a desire to hear how the artist would respond. As one viewer noted, "*It gave me an explanation, but it still felt like a starting point. I wanted to hear what the artist would actually say about it.*" (P37).

4.3 RQ3: Conditions and Effort for Artist Editing

4.3.1 Quantitative results. This analysis examined whether, within the observed workflow, human editing was associated with lower time cost than writing from scratch. Descriptive statistics showed that editing was generally faster than writing from scratch. This was consistent across all artists, whose individual means ranged from 62.8 to 73.7 seconds for editing and from 120.3 to 139.1 seconds for writing (Figure 8). As noted in Section 3.5.2, the observed differences are indicative and informative, though not conclusive.

To further examine what factors contribute to variation in editing effort, a multiple linear regression was conducted with editing time as the dependent variable and six predictors (see Table

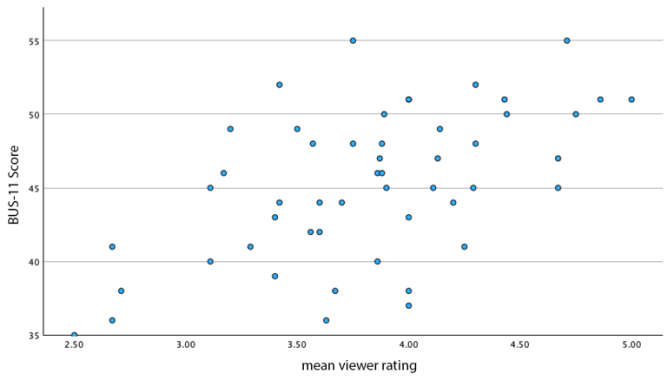


Figure 7: Association between mean viewer ratings and BUS-11 total scores.

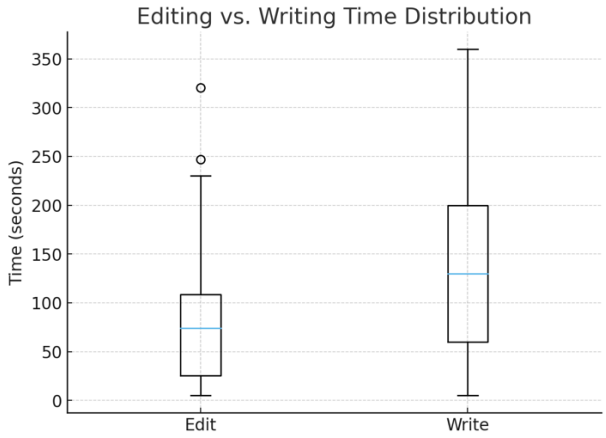


Figure 8: Comparison of editing time and writing-from-scratch time for artist responses.

2): artist_rating, viewer_rating, should_escalate, question_category, question_text_length, and artist_id. The overall model was significant ($F(6,229) = 7.55$, $p < .001$, $R^2 = .165$). Two predictors emerged as significant: artist_rating showed a clear negative effect ($\beta = -.535$, $p < .001$), indicating that higher-quality initial drafts required substantially less editing time, whereas viewer_rating showed a positive effect ($\beta = .619$, $p < .001$), indicating that higher viewer ratings were associated with longer editing times. Other predictors—including should_escalate, question_category, question_text_length and artist_id—were non-significant (all $p > .25$). Multicollinearity was not a concern (all VIF < 3.5). To rule out whether differences in baseline writing duration across artists contributed to variation in editing time, we fitted a regression model with write_time as the dependent variable and artist ID and artwork ID as categorical predictors. The overall model was not significant ($F(2, 233) = 0.279$, $p = .757$, $R^2 = .002$), indicating that neither artist ($p = .663$) nor artwork ($p = .468$) accounted for variation in writing duration. This result suggests that the observed variation in editing time was unlikely to be driven by systematic differences in baseline writing duration across artists.

4.3.2 Qualitative Results. Across all artist interviews, artists found that editing the chatbot-generated draft was substantially faster than composing an answer from scratch. They described the draft response as providing a reliable structural foundation; what one artist called "*already provides 70–80% of the skeleton I need*" (A2), which allowed them to focus on the parts of the response that required artist sensitivity rather than rebuilding the explanation entirely. Artists emphasised that the chatbot saved them from re-writing basic descriptions of techniques, materials, or compositional elements. As A1 noted, "*The chatbot draft already covers the fundamental explanation, so I don't need to write that again.*" Instead, the editing time was mainly spent on refining tone, adding contextual nuance, and articulating the emotional motivations behind a creative choice. One artist explained that their edits often aimed to "*add the feeling or the moment behind the painting, because that part can't come from the chatbot response*" (A3). However, artists also identified specific cases where editing took longer than writing, aligning with the quantitative findings. These were typical instances where the chatbot draft response introduced overly generic descriptions, unnecessary technical detail, or additional interpretations misaligned with the artist's main intent. As A1 described, "*Sometimes the draft is too long, or it explains too many details I prefer not to say. Then I have to clean up the draft response.*"

Multiple artists (A1, A2 and A4) pointed out that when the chatbot draft aligned well with their creative thinking, editing became a minor task. For example, A2 described: "*When the draft already matches my understanding of the work, I just need to polish a few sentences. It saves a lot of time.*" A4 held the same view, explaining that a good chatbot draft was one that "*responded carefully to the image and didn't invent things I never said*," reducing the need for structural rewriting. Artists also highlighted that even when chatbot draft responses were accurate, their richness or length sometimes required more careful refinement. As A3 described, "*Sometimes the chatbot provides a very good and detailed draft response. But I still need time to digest it, keep the parts that fit my style, and reorganise the flow.*" A1 similarly noted that, "*Detailed draft responses take longer to edit simply because I have to read through everything.*"

4.4 RQ4: Perceived Quality of Edited Responses

In the follow-up interviews with a subset of 12 viewers, interviewees reviewed the artist-edited responses corresponding to the IE questions they had asked during the gallery visit, together with the original chatbot draft responses. This allowed us to examine how these post-visit artist-edited answers were perceived. Through these interviews, participants collectively evaluated 59 artist-edited responses in relation to their corresponding chatbot drafts. Many of these responses (n = 39) were described as clearly better than the chatbot drafts, while a smaller number were perceived as more or less the same (n = 18). Only two responses were considered less helpful than the initial chatbot draft. Responses perceived as better were commonly described as richer, more expressive, and more personalised. Interviewees highlighted the artist's ability to introduce emotional context, personal experience, or intentional ambiguity that was absent from the chatbot's draft. One viewer remarked that the revised answer "*felt like a conversation—he talked about painting alone on a rainy night*" (P11). Another noted that "*the chatbot helped me understand the technique first, but the artist's reply explained why it mattered emotionally*" (P30). In cases where responses were perceived as similar, interviewees noted that the chatbot draft had already captured the essential descriptive or interpretive content, leaving limited room for further refinement. For example, one participant commented that "*the artist mostly rephrased what the chatbot had already said, so the meaning didn't change much*" (P42). The few responses perceived as less helpful were associated with instances where artists chose to substantially shorten the reply or remove speculative interpretations introduced by the chatbot. One interviewee explained that "*the edited version was more cautious, but I preferred the chatbot's longer explanation because it gave me more to think about*" (P9).

5 DISCUSSION

5.1 Revisiting Research Questions

5.1.1 RQ1: Routing Effectiveness and Boundary Cases. RQ1 examined how effectively the system's routing mechanism supported decisions about when viewer questions should be handled by the chatbot or escalated to the artist. Quantitatively, the classifier demonstrated good alignment with the artists' judgements, particularly in its ability to identify queries involving emotion, interpretation, or artistic intention. The relatively high recall is consistent with a conservative strategy that prioritised capturing potentially nuanced questions over the risk of failing to escalate them. Qualitative findings further underscored the perceived appropriateness of routing artist involvement for interpretive questions, reinforcing the content-based rationale underlying the system's escalation decisions.

However, the qualitative findings also revealed a nuance not reflected in the classifier metrics alone. A subset of questions occupied an ambiguous middle ground between factual clarification and interpretive inquiry, and these "in-between" cases did not elicit consistent expectations. Similar challenges around intent ambiguity and boundary cases have been documented in prior work on query classification and routing in hybrid human-AI systems [23, 53]. Follow-up interviewees and artists differed in whether escalation was necessary for similar queries, depending on whether a question was treated as straightforward clarification or as an invitation to explore deeper intent. Collectively, these patterns indicate that while the routing mechanism performed reliably in this study, determining whether a question warrants artistic input can vary depending on how its intent is interpreted. Rather than reflecting an irreducible subjectivity, this variability highlights a current limitation of ArtiQ, which was trained on a relatively small dataset, and suggests that broader training data may help mitigate such inconsistencies in future iterations.

5.1.2 RQ2: Viewer Evaluation of Immediate Chatbot Responses. RQ2 examined how viewers evaluated the immediate chatbot responses produced within the system's two routing pathways. Quantitatively, viewers gave higher immediate helpfulness ratings to chatbot-generated responses addressing AF questions than to chatbot-generated initial drafts addressing IE questions, and this pattern held across individual galleries. These differences were further supported by the positive correlation between per-answer ratings and BUS-11 scores, which in this study reflected viewers' overall satisfaction with the on-site chatbot interaction they experienced during the gallery visit.

Qualitatively, viewers described chatbot-generated responses to AF questions as fast, accurate, and helpful for clarifying techniques or drawing attention to details they had not previously noticed, suggesting that the chatbot's strength lies in providing informationally clear descriptions that support quick understanding. By contrast, viewers' comments on IE draft responses were less positive, with these responses more often treated as provisional starting points. Overall, the RQ2 findings suggest that such informational clarity in chatbot-generated responses supported viewers' satisfaction in the CO condition, whereas immediate responses in the CAY condition were more often experienced as partial answers that invited later artist refinement.

5.1.3 RQ3: Conditions and Effort for Artist Editing. RQ3 examined under what conditions human editing was justified within the hybrid workflow and what shaped the effort required from artists. Quantitatively, in this structured workflow, editing the chatbot's draft appeared to take less time than composing answers from scratch, suggesting that the drafts may help reduce some of the workload by providing an initial structure for artists to refine. The regression analysis further showed that variation in editing time was driven not by differences between artists, but by properties of the chatbot's draft response. In cases where viewers were already satisfied with the chatbot-generated draft response, additional artist editing may not always be necessary, suggesting opportunities for viewer-initiated stopping points that allow effective offloading of question-answering work to the chatbot. Such viewer feedback could be explicitly communicated to the artist as part of the workflow, allowing them to make a more informed decision about whether further editing effort is warranted. The qualitative findings deepen this picture (Section 4.3).

More broadly, these findings point to trust as a central factor in offloading creative question-answering work to AI. Prior work on draft-then-refine human-AI workflows has shown that AI-generated drafts can reduce expert workload when they are treated as provisional starting points and align with human expectations and domain intent [2, 26]. Our findings extend this line of work by

showing that, in interpretive art contexts, such alignment is closely tied to artists' trust in the chatbot's output: when artists regarded the drafts as consistent with their intentions, editing became a process of refinement rather than correction, enabling more substantial reductions in human workload. Conversely, when drafts were perceived as stylistically generic or misaligned, the efficiency benefits diminished, echoing concerns raised in prior studies about the cost of correcting low-quality or misaligned AI outputs [11].

5.1.4 RQ4: Perceived Quality of Artist-Edited Responses. RQ4 examined how viewers perceived the quality of artist-edited chatbot responses when these were presented to them after their gallery visit. Follow-up interviews revealed a complementary dynamic: when reviewing the artist-edited responses for their IE questions afterward, interviewees consistently highlighted the added emotional resonance and narrative depth that only the artists could provide. Several interviewees described the two stages as synergistic; the chatbot grounded their understanding, while the artist brought meaning and intention. However, this added value was not uniform across all cases: some edited responses were perceived as broadly similar to the draft, and a few were considered less helpful. Overall, the RQ4 findings suggest that, when presented after the gallery visit, artist refinement within the CAY workflow could often contribute to a deeper appreciation of the artworks. This pattern aligns with prior work in art and museum contexts showing that interpretive depth is more strongly associated with human expertise and authorial voice [12, 51].

5.2 Design Implications

Building on the empirical findings, our study suggests how an artist-in-the-loop chatbot like ArtiQ might be structured to better support art interpretation in gallery contexts. Accordingly, we outline three design implications that translate these insights into actionable guidance for future art chatbot systems in gallery contexts.

5.2.1 Designing Layered Response Structures. A layered design can mitigate two common risks when dealing with viewers' questions in a gallery context: chatbots overextending into speculative interpretation, and artists having to re-articulate foundational explanations that can be reliably automated. Treating the chatbot's contribution as an initial layer allows viewers to quickly orient themselves, while enabling artists to focus their effort on the aspects that require human sensitivity. This layered approach resonates with prior arguments that automated systems are well-suited for delivering objective or orienting information [38], while interpretive authority in cultural contexts remains closely tied to human expertise [19]. Interfaces may therefore present responses to AF questions as a primary layer, while reserving artist involvement for cases that are explicitly routed for interpretive refinement. In this sense, the chatbot should not be understood merely as a pipeline for delivering objective information, but as a mediating layer that helps viewers articulate their curiosity, prepares a shared point of reference for later artist involvement, and supports a more grounded transition from initial orientation to interpretive dialogue.

5.2.2 Routing for Ambiguity and Subjectivity. Our findings suggest that whether artist intervention is warranted cannot be determined solely by system prediction or artist preference; it must also remain responsive to what viewers are actually seeking from the interaction. Rather than enforcing a rigid binary boundary, future systems should introduce a third handling mode specifically for "in-between" questions. This extends prior work on intent-aware routing, which has primarily focused on technical accuracy, by highlighting the need to account for subjectivity and interpretive variability in cultural domains [23, 53]. When the system detects that a query could plausibly be interpreted in more than one way, it can surface this ambiguity to the viewer, providing a concise chatbot response while also offering the option to request the artist's perspective. This viewer-mediated routing allows individuals to decide whether they seek a straightforward explanation or further insight, accommodating the variability observed in the preferences of both viewers and artists.

In parallel, artists could annotate such ambiguous cases within their interface (e.g., indicating whether they believe similar questions are better directly handled by the chatbot or by forwarding to themselves). These annotations could be fed back into the system over time, enabling the classifier to iteratively refine how it distinguishes between questions that warrant artist involvement and those that can be handled by the chatbot alone, gradually aligning its routing behaviour with individual artists' interpretive styles and thresholds for intervention. Over time, this artist-informed learning process may support more consistent and trusted routing decisions, helping to support artist-in-the-loop workflows at scale without requiring continuous manual oversight.

5.2.3 Optimising chatbot drafts for efficient and trustworthy human refinement. Building directly on the RQ3 finding that variation in editing effort was closely associated with properties of the chatbot's drafts, future art chatbot systems should intentionally shape draft responses for editability and alignment, rather than for maximal completeness: providing clear structural scaffolding, accurate grounding, and explicit alignment with the artist's stance. In practice, such a structure can be shaped at the generation level through prompt design and output constraints, for example, by explicitly requesting concise, sectioned responses that prioritise core factual grounding while deferring speculative or emotional interpretation. Because overly long or exhaustive drafts can increase editing effort by requiring artists to remove or reframe content rather than refine it, controlling response length and scope through prompts represents a practical design lever for balancing informativeness with editability. This observation complements prior findings that overly verbose or misaligned AI drafts can increase revision burden, particularly in expert-driven workflows where correction costs outweigh generative benefits [11, 26].

Importantly, informativeness need not be sacrificed. Supporting optional viewer-initiated stopping points, where the interaction can conclude without escalation, creates opportunities to offload satisfactory question-answering work to the chatbot while preserving artist effort for cases where personal perspective is most needed. For cases that do require refinement, draft structure remains critical: drafts that separate core factual grounding from optional detail, visually differentiate response layers, and flag uncertain elements enable artists to edit efficiently, echoing artists' descriptions of drafts as useful scaffolding rather than finished answers.

5.3 Practical Implications

Beyond interaction-level design considerations, our findings also carry practical implications for deploying ArtiQ in real-life gallery settings. Its hybrid temporal structure—immediate chatbot responses on site, asynchronous artist-edited responses later—allows engagement to continue beyond a single encounter, combining on-site interaction with post-visit reflection. At the same time, asynchronous responses raise practical considerations related to viewers' attention and memory: if artist feedback arrives too late, viewers may lose interest or forget the context of their original question. Rather than prescribing a fixed "optimal" waiting time, our findings point to the importance of expectation management and continuity mechanisms. Some of these are already implemented in ArtiQ, such as making response status visible and allowing viewers to revisit their previously submitted questions; however, these features were not systematically evaluated in the present study, and their effectiveness cannot be assumed. For example, indicating that an artist response will arrive at an unspecified later time may risk frustrating viewers if delays are long or unpredictable.

From the artists' perspective, asynchronous response workflows are equally critical. Artist-edited replies in ArtiQ are not intended to function as real-time obligations, as imposing strict deadlines would risk shifting the system from supportive to burdensome. This asynchronous structure is therefore a deliberate design decision to preserve artists' autonomy over when and how they engage, respecting the natural rhythm of artistic practice rather than accelerating it. The reductions in editing time observed in RQ3 should thus be understood not as a pressure to respond faster, but as a means of making participation sustainable when artists choose to engage.

At an institutional level, ArtiQ also presents both opportunities and risks for galleries. One clear benefit is its potential to sustain relationships between viewers and artists beyond physical exhibitions, transforming one-time encounters into ongoing dialogue: follow-up responses can reinvigorate interest in artworks after a visit and encourage return engagement. This suggests a broader role for AI in art appreciation—not only answering isolated questions, but facilitating a more dialogic exchange in which viewers request clarification, artists selectively extend or qualify prior answers, and understanding develops across multiple turns rather than through a single response. At the same time, galleries must consider the additional demands placed on artists: while articulating perspectives on existing works may stimulate reflection and even inspire new creative directions, excessive or poorly managed volumes of questions risk diverting time away from artistic production. These tensions underscore why the efficiency and selectivity mechanisms discussed earlier are not merely technical refinements, but essential safeguards for long-term viability.

5.4 Limitations and Future Research

Consistent with many empirical studies conducted under resource constraints, this work has some limitations that are important to acknowledge. First, the study was situated within four gallery sites specialising in traditional Chinese painting, involving four artists and 51 viewers. While this design offered strong ecological validity and a coherent domain for examining human-AI collaboration, it also limits the generalisability of our findings to other artistic genres or culturally diverse viewers. This constraint reflects an intentional methodological choice: restricting the artistic domain allowed us to maintain a consistent knowledge base and reduce confounds introduced by cross-genre variation. Future research should expand this work by deploying the system across more varied artistic media (e.g., oil painting, sculpture, photography), and cross-cultural contexts. Such studies would allow systematic examination of whether the layered response structure and routing dynamics identified here hold under different interpretive traditions and viewer expectations.

Second, the effectiveness of the system's two-step response process was influenced by the quality of the chatbot-generated draft responses, which were produced using a single LLM configuration. This choice allowed us to isolate how artists interacted with a consistent draft structure, but it also means that our findings reflect the behaviour of one particular model rather than a broader range of generative approaches. Future studies could examine how different LLMs affect draft usefulness, editing effort, and alignment with artists' preferences.

Third, a methodological limitation is that on-site viewer ratings captured only the immediate chatbot-response stage rather than the full artist-refined experience. Although asynchronous artist replies are part of the intended real-world deployment model of ArtiQ, most viewers in the present study did not receive or evaluate those artist-edited responses. As a result, the study does not provide a full evaluation of the complete layered interaction experience, and follow-up interviews with a subset of viewers cannot fully substitute for the more natural conditions under which delayed artist-edited responses would be encountered. Future work should therefore examine how delayed artist responses are received under more complete deployment conditions, including how memory, context, and post-visit reflection shape viewer experience over time.

Fourth, this study did not examine issues of emotional engagement or trust, as these were beyond its defined scope and were excluded to maintain focus. One possible factor influencing viewer and artist satisfaction with chatbot responses is whether the response's affective or emotional tone aligns with that of the question. For AF questions, such resonance may be high, whereas for IE questions, it is likely to be less predictable. Additionally, the complex relationship between emotion and trust warrants further investigation. Previous research [17] suggests that positive emotions are associated with higher trust, which in turn can enhance user experience and satisfaction. We plan to address this aspect explicitly in future work, for example by asking viewers and artists to indicate their trust levels and emotional responses during interactions with ArtiQ.

6 CONCLUSION

This paper examined how artist-in-the-loop workflows can support diverse viewer questions in art galleries, where informational clarity and interpretive depth must be carefully balanced. Through the design and multi-site field deployment of ArtiQ, we investigated

how routing mechanisms, layered responses, and draft-based collaboration structure interactions between viewers, chatbots, and artists in real-life exhibition settings.

Our findings provide empirical insights into the complementary roles of chatbots and human artists in art interpretation. Chatbot-generated responses effectively supported viewers' immediate orientation and factual understanding, while artist involvement contributed interpretive nuance, emotional resonance, and contextual grounding that viewers consistently valued. Importantly, the draft-then-refine workflow could reduce artists' effort compared to composing responses from scratch, suggesting that human authorship can be preserved while making participation more sustainable. At the same time, the boundary cases observed in routing decisions highlight the inherently subjective nature of art-related questioning and suggest that interpretive need is not always clear-cut.

This work contributes a field study of artist-in-the-loop interaction in a cultural context, extending prior research on hybrid human-AI workflows into the domain of art appreciation. Rather than positioning chatbots as substitutes for human interpretation, our findings highlight their role as infrastructural supports that scaffold understanding while leaving interpretive authority with human experts. In this sense, ArtiQ should be understood not as a fully resolved end-to-end solution, but as empirical support for a layered workflow that is both promising and worth developing further. While grounded in traditional Chinese ink painting, this study sets the stage for future research to examine how artist-in-the-loop approaches may be adapted to other artistic domains and cultural settings. In the long term, ArtiQ has the potential to significantly impact art education and help preserve cultural heritage by sustaining curiosity and fostering appreciation of traditional Chinese ink painting.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Olatundun D. Awosanya, Alexander Harris, Amy Creecy, Xian Qiao, Angela J. Toepp, Thomas McCune, Melissa A. Kacena, and Marie V. Ozanne. 2024. The Utility of AI in Writing a Scientific Review Article on the Impacts of COVID-19 on Musculoskeletal Health. Current Osteoporosis Reports 22, 1: 146–151. https://doi.org/10.1007/s11914-023-00855-x

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

[4] Simone Borsci, Martin Schmettow, Alessio Malizia, Alan Chamberlain, and Frank Van Der Velde. 2023. A confirmatory factorial analysis of the Chatbot Usability Scale: a multilanguage validation. Personal and Ubiquitous Computing 27, 2: 317–330. https://doi.org/10.1007/s00779-022-01690-0

[5] Simone Borsci and Martin Schmettow. 2024. Re-examining the chatBot Usability Scale (BUS-11) to assess user experience with customer relationship management chatbots. Personal and Ubiquitous Computing 28, 6: 1033–1044. https://doi.org/10.1007/s00779-024-01834-4

[6] Susan Bush and Hsio-yen Shih (eds.). 2012. Early Chinese texts on painting. Hong Kong University Press, Hong Kong.

[7] James Cahill (ed.). 1977. Chinese painting. Rizzoli Internat. Publ, New York, NY.

[8] Francesco Colace, Marco Lombardi, and Domenico Santaniello. 2020. A Chatbot for supporting users in Cultural Heritage contexts (S). 22–27. https://doi.org/10.18293/DMSVIVA20-009

[9] Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A Complete Survey on LLM-based AI Chatbots. https://doi.org/10.48550/ARXIV.2406.16937

[10] Kohinoor M. Darda and Anjan Chatterjee. 2023. The impact of contextual information on aesthetic engagement of artworks. Scientific Reports 13, 1: 4273. https://doi.org/10.1038/s41598-023-30768-9

[11] Wafeequa Dinath, Sithembiso Khumalo, and Bonolo Mashigo. 2025. Evaluating the Effectiveness of a Rule-Based WhatsApp Chatbot in Automated Query Resolution. European Conference on Innovation and Entrepreneurship 20, 1: 208–216. https://doi.org/10.34190/ecie.20.1.3838

[12] Olga Dysthe. 2021. Opportunities and challenges of dialogic pedagogy in art museum education. Dialogic Pedagogy: An International Online Journal 9: A1–A36. https://doi.org/10.5195/dpj.2021.317

[13] Vicente Estrada-Gonzalez, Scott East, Michael Garbutt, and Branka Spehar. 2020. Viewing Art in Different Contexts. Frontiers in Psychology 11: 569. https://doi.org/10.3389/fpsyg.2020.00569

[14] John H Falk and Lynn D Dierking. 2016. The Museum Experience Revisited. Routledge. https://doi.org/10.4324/9781315417851

[15] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. Interactions 24, 4: 38–42. https://doi.org/10.1145/3085558

[16] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In Internet Science, Svetlana S. Bodrunova (ed.). Springer International Publishing, Cham, 194–208. https://doi.org/10.1007/978-3-030-01437-7_16

[17] Asbjørn Følstad, Effie L.-C. Law, and Nena Van As. 2024. Conversational Breakdown in a Customer Service Chatbot: Impact of Task Order and Criticality on User Trust and Emotion. ACM Transactions on Computer-Human Interaction 31, 5: 1–52. https://doi.org/10.1145/3690383

[18] Ana Isabel González-Herrera, Andrea Betsabé Díaz-Herrera, Paula Hernández-Dionis, and David Pérez-Jorge. 2023. Educational and accessible museums and cultural spaces. Humanities and Social Sciences Communications 10, 1: 67. https://doi.org/10.1057/s41599-023-01563-8

[19] Giuliana Guazzaroni. 2021. Digital Heritage: New Ways to Provoke an Emotional Response to Art. International Journal of Art, Culture, Design, and Technology 10, 1: 1–17. https://doi.org/10.4018/IJACDT.2021010101

[20] David Henry. 2015. Talking Deeper about Cultural Difference: A Digital Interactive from Melbourne. Curator: The Museum Journal 58, 2: 209–222. https://doi.org/10.1111/cura.12108

[21] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1: 1–20. https://doi.org/10.1145/3579628

[22] Housen, Abigail. 2007. Art Viewing and Aesthetic Development: Designing for the Viewer. In From Periphery to Center: Art Museum Education in the 21st Century, Pat Villeneuve (Ed.). National Art Education Association, Reston, VA, 172–189.

[23] Oksana Ivashchenko, Stanislav Filip, and Bohdan Ratushnyi. 2025. DEVELOPMENT OF AN EDUCATIONAL CHATBOT WITH A CONTEXTUAL INTENT SYSTEM ON THE DIALOGFLOW PLATFORM. Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies, 1 (13): 17–24. https://doi.org/10.20998/2079-0023.2025.01.03

[24] Juan Carlos Jaramillo, Daniele Occhiuto, and Franca Garzotto. 2018. Artworks' Features Discovery Through Engaging Conversations for Children. IOP Conference Series: Materials Science and Engineering 364: 012097. https://doi.org/10.1088/1757-899X/364/1/012097

[25] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons 61, 4: 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

[26] Melissa A. Kacena, Lilian I. Plotkin, and Jill C. Fehrenbacher. 2024. The Use of Artificial Intelligence in Writing Scientific Review Articles. Current Osteoporosis Reports 22, 1: 115–121. https://doi.org/10.1007/s11914-023-00852-0

[27] Juyeon Kim, MyoungHun Han, SeungJun Kim, and Jin-Hyuk Hong. 2026. Human-AI Collaboration in Generating Graphical Museum Descriptions. Journal on Computing and Cultural Heritage 19, 2: 1–23. https://doi.org/10.1145/3786327

[28] Florian Leiser, Sven Eckhardt, Merlin Knaeble, Alexander Maedche, Gerhard Schwabe, and Ali Sunyaev. 2023. From ChatGPT to FactGPT: A Participatory Design Study to Mitigate the Effects of Large Language Model Hallucinations on Users. In Mensch und Computer 2023, 81–90. https://doi.org/10.1145/3603555.3603565

[29] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. 2013. UMUX-LITE: when there's no time for the SUS. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2099–2102. https://doi.org/10.1145/2470654.2481287

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 174, 9459–9474.

[31] Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren Wei. 2024. Ethical Considerations and Fundamental Principles of Large Language Models in Medical Education: Viewpoint. Journal of Medical Internet Research 26: e60083. https://doi.org/10.2196/60083

[32] Yannan Lin. 2025. The Role of Art Audience in the Changing Times: From Passive Acceptance to Participation. Lecture Notes in Education Psychology and Public Media 72, 1: 38–42. https://doi.org/10.54254/2753-7048/2025.ND19109

[33] Jinyu Liu, Effie Lai-Chong Law, and Hubert P. H. Shum. 2025. Exploring Artists' and Art Viewers' Perspectives for Art Chatbots: Implications for a Design Framework. In Proceedings of the 7th ACM Conference on Conversational User Interfaces, 1–17. https://doi.org/10.1145/3719160.3736626

[34] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 5286–5297. https://doi.org/10.1145/2858036.2858288

[35] Michael McTear. 2021. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-031-02176-3

[36] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. In Designing Interactive Systems Conference 2021, 926–941. https://doi.org/10.1145/3461778.3462143

[37] Davide Picca. 2024. Emotional Hermeneutics. Exploring the Limits of Artificial Intelligence from a Diltheyan Perspective. In Proceedings of the 35th ACM Conference on Hypertext and Social Media, 12–16. https://doi.org/10.1145/3648188.3680255

[38] Sajjadur Rahman and Eser Kandogan. 2022. Characterizing Practices, Limitations, and Opportunities Related to Text Information Extraction Workflows: A Human-in-the-loop Perspective. In CHI Conference on Human Factors in Computing Systems, 1–15. https://doi.org/10.1145/3491102.3502068

[39] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. In Designing Interactive Systems Conference, 757–781. https://doi.org/10.1145/3643834.3660729

[40] Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering 19, 1: 10–26. https://doi.org/10.1631/FITEE.1700826

[41] Adam Sobieszek and Tadeusz Price. 2022. Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. Minds and Machines 32, 2: 341–364. https://doi.org/10.1007/s11023-022-09602-0

[42] Branden R. Sosa, Michelle Cung, Vincentius J. Suhardi, Kyle Morse, Andrew Thomson, He S. Yang, Sravisht Iyer, and Matthew B. Greenblatt. 2024. Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries. Journal of Orthopaedic Research 42, 6: 1276–1282. https://doi.org/10.1002/jor.25782

[43] Eva Specker, Pablo P. L. Tinio, and Michiel Van Elk. 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, 3: 265–275. https://doi.org/10.1037/aca0000107

[44] Philipp Spitzer, Joshua Holstein, Patrick Hemmer, Michael Vössing, Niklas Kühl, Dominik Martin, and Gerhard Satzger. 2025. Human Delegation Behavior in Human-AI Collaboration: The Effect of Contextual Information. Proceedings of the ACM on Human-Computer Interaction 9, 2: 1–28. https://doi.org/10.1145/3710999

[45] Kamila Štekerová. 2022. Chatbots in Museums: Is Visitor Experience Measured? Czech Journal of Tourism 11, 1–2: 14–31. https://doi.org/10.2478/cjot-2022-0002

[46] Eirini Eleni Tsiropoulou, Athina Thanou, and Symeon Papavassiliou. 2017. Quality of Experience-based museum touring: a human in the loop approach. Social Network Analysis and Mining 7, 1: 33. https://doi.org/10.1007/s13278-017-0453-2

[47] Savvas Varitimiadis, Konstantinos Kotis, Andreas Skamagis, Alexandros Tzortzakakis, George Tsekouras, and Dimitris Spiliotopoulos. 2020. Towards implementing an AI chatbot platform for museums. International Conference on Cultural Informatics, Communication & Media Studies 1, 1. https://doi.org/10.12681/cicms.2732

[48] Savvas Varitimiadis, Konstantinos Kotis, Dimitra Pittou, and Georgios Konstantakis. 2021. Graph-Based Conversational AI: Towards a Distributed and Collaborative Multi-Chatbot Approach for Museums. Applied Sciences 11, 19: 9160. https://doi.org/10.3390/app11199160

[49] Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. 2023. A Survey of the Evolution of Language Model-Based Dialogue Systems: Data, Task and Models. https://doi.org/10.48550/ARXIV.2311.16789

[50] Huan Wang and Andrii Matviienko. 2025. Experiencing Art Museum with a Generative Artificial Intelligence Chatbot. In Proceedings of the 2025 ACM International Conference on Interactive Media Experiences, 430–436. https://doi.org/10.1145/3706370.3731650

[51] Budi Wibawa and Zulidyana Dwi Rusnalasari. 2024. Navigating Ethics and Innovation: The Role of AI in Cultural Heritage. Harmonia: Journal of Music and Arts 2, 4: 198–212. https://doi.org/10.61978/harmonia.v2i4.905

[52] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems 135: 364–381. https://doi.org/10.1016/j.future.2022.05.014

[53] Ghanshyam S. Yadav, Kshitij Pandit, Erin A. Gross, Karandeep Singh, D. Yvette LaCoursiere, Ming Tai-Seale, Marlene Millen, Cynthia Gyamfi-Bannerman, Sarah Parker, Divya Gupta, and Christopher A. Longhurst. 2025. Evaluating Generative Artificial Intelligence-Drafted Responses for Patient Messages in Obstetrics and Gynecology. O&G Open 2, 5. https://doi.org/10.1097/og9.0000000000000131

[54] Zihan Yu. 2025. Beyond the Algorithm: How AI is Reshaping Museums, from Preservation to Participation. AI & Future Society 1, 1: 6–12. https://doi.org/10.63802/afs.v1.i1.86

[55] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In Proceedings of the CHI Conference on Human Factors in Computing Systems, 1–18. https://doi.org/10.1145/3613904.3642343

[56] Shuning Zhang, Hui Wang, and Xin Yi. 2025. Exploring Collaboration Patterns and Strategies in Human-AI Co-creation through the Lens of Agency: A Scoping Review of the Top-tier HCI Literature. Proceedings of the ACM on Human-Computer Interaction 9, 7: 1–43. https://doi.org/10.1145/3757594