Conversation agent leveraging VTS for opinion forming during an art exhibition

DANIEL STANCHEV, University of Twente, The Netherlands

This paper combines two fields of study: conversational artificial intelligence and visual processing techniques. Chatbots are software that uses natural language understanding put in context to converse with users. Visual Thinking Strategies (VTS) describes a methodology intended to improve discussion and observation skills in novice art observers. This paper aims to develop a conversation agent leveraging Large-Language Models (LLMs) and VTS to gather insightful knowledge about user interpretation of an artwork while retaining optimal satisfaction. This work extends pre-existing research of audio commentary analysis during art viewing and adds to the existing body of knowledge regarding VTS, communicative AI, and art appreciation techniques.

Additional Key Words and Phrases: VTS, LLMs, user satisfaction, analysis, art appreciation technique, opinion forming, artwork, visual processing technique, artificial intelligence, chatbot, Natural Language Processing, Natural Language Generation, knowledge graph,

1 INTRODUCTION

With the introduction of Large-Language Models (LLMs), chatbots have become much more common in many industries[17]. Service provider websites utilize artificial assistance, leveraging the near-constant availability, vast knowledge, and patience to help as many users as possible. Despite this, there is an area that has not utilized chatbots all that much, namely art exhibition establishments such as museums and galleries. There are examples [27, 31] that focus mainly on answering user inquiries, usually educational answers with factual information about the piece.

A more engaging approach involves encouraging the user to carefully observe and express their thoughts. This methods allows viewers to engage with the art on a much deeper level fostering in-depth understanding and appreciation for the artwork. Such engagement helps form long term memories and knowledge of the piece [16]. To achieve this we propose using techniques that improve observation skills through a guided discussion. Specifically, we will aim to develop a conversational agent designed to adapt Visual Thinking Strategies (VTS). This agent will aim to improve the user's ability to form and express opinions by guiding them in observing and discussing art.

2 PROBLEM STATEMENT

Constructing a clear and concise opinion on an art piece can be difficult, as most people do not have the tools in the form of visual literacy or the interest to take the time and really examine an artwork [11]. To mitigate these issues, we will use Visual Thinking Strategies [33], due to the extensive knowledge and research showing positive impact in visual and critical thinking skills. Despite this, there is a lack of work striving to integrate it with technology, such as Large Language Models (LLMs). The technology powering the recent advancements in generative artificial intelligence, allowing for the generation of naturally spoken language [20]. This paper will aim to address this gap. To achieve this, we will incorporate these strategies into a chatbot, whose goal is fostering a discussion and gathering knowledge of user interpretations while viewing the art.

2.1 Research Question

The problem will be addressed by the following research question:

How can a conversation agent leveraging Visual Thinking Strategies (VTS) be used into an art exhibition to increase observation and opinion forming skills while retaining user satisfaction?

Note, that user satisfaction here refers to avoiding frustrating the user to the point where they do not see a point in engaging with the agent. We will attempt to answer the main question with the following sub-questions:

- How can we translate practices from VTS to a custom chatbot leveraging LLMs?
- What methods can be employed to make the agent's responses more focused and context aware through the conversation so as to not deter from the user experience?

3 RELATED WORK

Materials were gathered from Scopus, IEEE, ACM, and Semantic Scholar with the following search terms "visual thinking", "visual thinking strategies", "visual thinking in the museum", VTS, and VTM, to find papers regarding Visual Thinking Strategies and other techniques utilized to improve art viewing. Separately, we looked for papers with the keywords rasa, chatbot, artificial intelligence, ask, question, knowledge, and extraction to look into chatbots and other systems that attempted to make task-oriented chatbots. Additionally, the tool Elicit [7] was used to search for materials on testing and evaluation of chatbots, and tour guides. The inclusion criteria for considering works were being sourced from a scientific database and having a minimum of one citation, resulting in 22 works spread into three categories relevant to this research.

3.1 Visual Thinking Strategies

Abigail Housen and Philip Yenawine collaborated over ten years to create a methodology that improves visual observation skills, namely Visual Thinking Strategies. In "Essentials of Teaching and Integrating Visual and Media Literacy" a chapter on VTS [11] describes five aesthetic stages of visual literacy with most viewers being in the bottom two stages. Stage 1 is limited to short basic observations based only on personal experience. The second stage

TScIT 41, July 5, 2024, Enschede, The Netherlands

is where basic visual literacy is achieved. At this stage, viewers compare aspects of the artwork to their perception, knowledge of the world, and values. The result of acquiring better observation skills [13] is more time spent observing the artwork and having more to say when describing a piece. The book also claims that there are benefits to a range of skills not confined to the field of art [5, 9, 30], such as critical thinking, consistently backing claims with evidence, and considering multiple interpretations. In stage 3 the history, time frame, school, and other facets of art creation are considered. Stage 4 is where feelings and intuition enter, allowing the user to think of the meaning and symbolism behind object and people featured in the work. The final stage describes viewers who have spent a long time appreciating art to the point where they develop a more intimate connection to it, similar to that of an old friend. This stage is unfortunately, not a feasible goal for this research and will not be considered. To summarize, VTS is a method that allows for the formation of more detailed and authentic opinions while also improving other non-art-related skills, and this paper will be focusing on the first four stages of it.

3.2 Conversational Agent

As for the chatbot, we will use the RASA conversational agent software [3]. This framework allows for the creation of custom assistants with specified goals. Multiple papers utilize RASA to make such assistants in fields ranging from tutoring [28] and medicine [10] to finance [14]. There is also an example of a museum tour guide [27]. While most of these examples use RASA to answer user queries, it can also serve for information gathering. Additionally, the API used for response generation can function well on its own, as there are examples showcasing its ability to go into different roles like a support agent [32] and medical advisor [6, 26] successfully. Finally, there is a paper [16] that outlined the importance of momentary success for the accumulation of interest.

3.3 Testing Measurements

Regarding the formulation of testing measures, which aim to evaluate the prototype's effectiveness in enhancing user opinion formation, its ability to engage in natural conversation without causing frustration, and its adaptation of VTS, I collected eleven papers. The most influential among them was a work regarding the evaluation of chatbots in healthcare [1].This study outlined and categorized metrics that were used in over 60 other studies. Despite the fact that my research differs in field from the healthcare study, both share a goal-oriented approach aim to minimize user frustration while interacting with an agent. Therefore, the metrics identified in the healthcare study are relevant and applicable to my research as well. Similarly, there was another study [4] that collected and analysed evaluation methodologies of chatbots in different fields from 2016-2020 to identify common trends. The metrics in that paper were classified based on perspective, which narrowed down the criteria for selecting relevant metrics to the current study. A different work [18] surveyed artificial assistants with emphasis on fewer evaluation metrics in greater detail of which relevant to this research are the two automatic methods for measuring goal-oriented bots BLEU[23] and ROUGE[15]. Moreover, a paper [19] on the topic of paraphrasing used an evaluation method for the appropriateness of its responses. Furthermore, there was an article [29] that found users favor using such assistants for answering FAQ questions. In addition, two works reported on the performance of guides and their importance for the enjoyment and memory persistence of tour participants. A paper that investigated performance from the perspective of the viewer [2] concluded a combination of professional, personal, and social skills along with, personality are crucial to a guide's success. Another work outlined the importance of building a connection and rapport with a guide to be an essential part of tour enjoyment [12]. Finally, there were two articles [8, 28] that investigate the optimal number of participants for controlled experiments and usability tests.

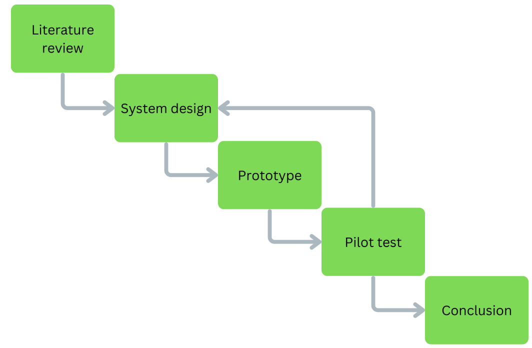

Fig. 1. Study design

4 METHODS OF RESEARCH

This section presents the procedure in which this study was carried out (see Figure 1). First, a literature review was conducted on three topics VTS, chatbots in different roles, and evaluation of chatbots. Then, a list of requirements (see Appendix 1) and a conversation flow (Figure 2) were made. The requirements are based on research of similar projects and the conversation flow went through multiple iterations initially inspired by a chapter from the book Essentials of Teaching and Integrating Visual and Media Literacy [11] and later improved by a pilot test.

4.1 Conversation strategy

To answer the first sub-question, we outline four stages implemented as the following rasa flows: Observation, Describe story, Describe author technique, and Describe interpretation. For each section, the chatbot will prompt the user with context-specific questions with adherents to the three research-tested questions from the book [11], for example: "What do you see?" for Observation and respond appropriately with commendations and remarks. To see an example of what the conversation might be like refer to Appendix 3. Note that the example conversation is an ideal scenario where none of the bugs or design flaws were triggered.

4.1.1 Stage 1: Observation. Most closely related to the first stage of aesthetic development here, the agent will be asking general questions about what is physically in the painting. An example question for this stage would be, "What do you see in the artwork?". This step aims to create a starting point, which the future stages

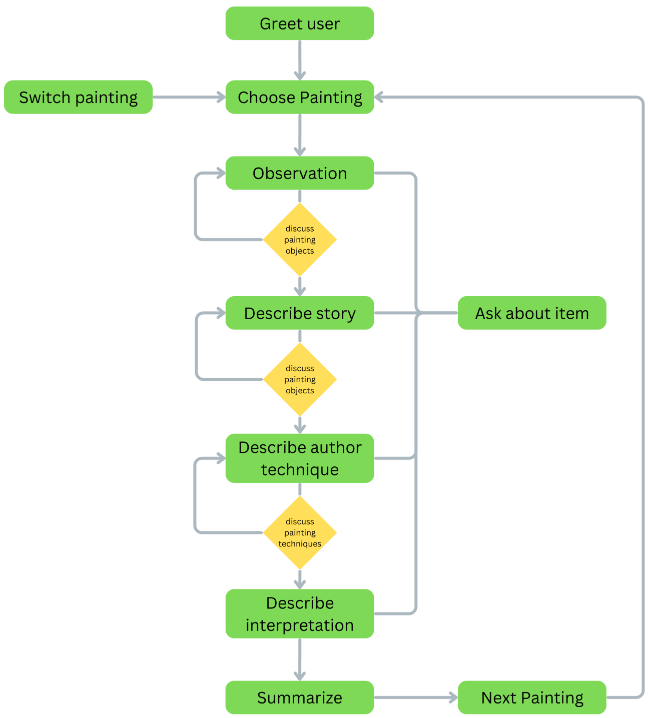

Fig. 2. Conversation flow

can use to build upon. The bot uses a database to determine what is in the artwork and prompts the user with an object if they are struggling. Struggling is considered as the user asking for the agent's opinion or help in identifying objects. We move on to the next stage after the conversation covers at least 50% of the known items. This percentage was chosen for two reasons, the user might not want to talk about all of the objects in the painting, and two the function that detects said objects is not always accurate and would sometimes overlook some items. Additionally, some paintings have very few known items stored in the database, which makes it more reliant on the detection function. The 50% coverage makes it easier to move on to other stages and avoid the possibility of the conversation getting stuck in a single stage.

4.1.2 Stage 2: Describe story. In stage two, the user will be prompted to create a narrative for the art piece to make them look closer and connect some of the observed objects from the previous stage. The goal here is to encourage them to be more constructive towards the piece by using their perception, as described in the second stage of aesthetic development. A question from this stage would be "How do you think <item> adds to the story in the painting?". Identical to the previous stage, the agent will decide to transition once the conversation has covered at least 50% of the known items in the painting for the same reasons as discussed in section 4.1.1.

4.1.3 Stage 3: Describe author technique. Stage three explores the techniques used to make the artwork, to educate the user, and to make them think about the author's intent. This block relates to stage three of aesthetic development, which has much fewer members, as it is harder to achieve due to the requirement of art knowledge that most regular museum goers do not have. An example question would be "What do you notice about the way the artist applied the paint in this artwork?". The agent will be more descriptive here to prevent the user from feeling uncomfortable due to their lack of knowledge by asking more focused questions, adding the name of a known material or technique, and asking the user what they think about it. Similarly to the last two sections, the bot will have access to a list of genres and materials to help the user if they are struggling and remember what is true regarding the artwork. The definition of struggle here is the same as in stage 1. The trigger to move on to the next stage is again covering over 50% of the genres and materials for the same reason as discussed in section 4.1.1.

4.1.4 Stage 4: Describe interpretation. Here, the bot will ask the user to describe their interpretation concerning symbolism and emotions portrayed by the artwork. This step tries to adapt the fourth stage of aesthetic development, which is even harder to reach as it requires lots of practice in art viewing. Therefore, to allow less experienced users to participate in the discussion, the bot may use interpretations stored in a knowledge graph to assist the user. Here is an example question "How does the portrayal of the man in the painting as a king influence your overall interpretation of the artwork?". Because interpreting art is highly subjective, the examples will be taken from previous interactions with past viewers. As an example consider "the expensive clothes and golden jar portray a feeling of importance and grandeur". At the end of the discussion, the bot will store the interpretation of the current viewer in the knowledge graph by adding it to a property list called interpretations associated with a painting. The conversation moves on from this stage once the user declares they have nothing else to say for example by typing "I cannot think of any other interpretation".

4.1.5 Conversation repair. To ensure the conversation remains natural throughout the entire interaction, the block "ask about item" serves to answer user queries regarding the artwork. That is done through the rasa component FlowPolicy which supports out-of-scope recognition and response, signifying to the user that they are going off-topic and the conversational agent can not assist them, for example by asking for the weather. The in-built rasa responses were altered to lead to custom actions in the case of the ask about item block and another was modified to stop the discussion after confirming with the user. Finally, viewers are free to switch to a different painting in any part of the discussion.

After going through all stages, the bot will summarize the conversation and include the main takeaways about the painting. To do so, the model considers the painting's description and conversation history. This step aims to show the overall learning goals in a digestible format to foster long-term recollection of the findings.

4.2 Conversation awareness and barriers

To ensure the agent remains focused and context aware three distinct components were utilized: specific instructions, a relevant history log, and knowledge graph information. The three parts were used in the prompt formation (see Appendix 2) of the API call.

4.2.1 Instructions. The instructions describe the role and competencies of the assistant, the environment that the interaction happens in and rules of the conversation based on stage and purpose. In the process of constructing the rules the system makes a database call to a knowledge graph which is addressed in section 4.2.3. From the example the instruction section is defined by the initial 15 sentences.

4.2.2 Relevant history. The relevant history log takes all messages from the rasa's conversation history object and includes only the entries of the user and assistant, that are relevant to the current painting. This is done by going through all events in the conversation history and taking only the ones from the assigning of the current painting to choosing a new one. In the example the relevant history is defined by the messages with roles assistant or user,

4.2.3 Knowledge graph information. The knowledge graph information is either a short painting description or a list of known items depending on the current conversation stage. The graph contains information on over 2990 painting of which we use 4. Each painting is saved as a node that has relationships with nodes in different categories Exhibition, Collection, Person, Item, Detail, Genre, Material, Keyword, and Content. Note that not all paintings have connections to nodes in each category. Additionally, each node has some properties attached to them, painting has a description for example. This data allows the chatbot to keep track of what is in the painting and prevent getting deceived by the user be it on purpose or by accident. Objects are stored and extracted from the knowledge graph based on the stage and active painting. For example the prompt for the observation/describe story stage shown in Appendix 2 uses nodes with the Item and Content roles. The describe author technique extracts the Genre and Materials nodes and the describe interpretation considers only the interpretation property assigned to the painting. Finally, the summary stage collects only the description associated with each painting.

4.3 Process

We used rasa flows to implement the design mentioned above. Starting with an introductory flow that greets the user, followed by one that handles painting selection, and then we have the blocks that implement VTS as described above, the summary part, and finally next painting, responsible for clearing saved information to prepare for the next round of discussion. Additionally, there's the 'ask about item' feature to address cases where users inquire about specific elements in the painting. There's also the 'switch painting' functionality, which allows for changing the artwork mid-discussion. Note that such changes restart the discussion from the initial topic, i.e., observation.

In the next step, we used the OpenAI's [22] LLM model to generate responses for each stage. This is done by sending a prompt to the OpenAI API. The prompt consists of a list of messages, the first being the instruction mentioned in section 4.2.3 with the role system. Followed by the relevant history of the conversation, with all messages having the role of either user or assistant. For an example please refer to Appendix 2, which shows the instructions prompt for the observation and describe story stage.

To ensure the model is aware of what is in the painting when constructing the prompt, we utilized a Neo4j [21] knowledge graph. We used 4 paintings from that knowledge graph, King Caspar, Head of a Boy in a Turban, Diego Bemba, and Pedro Sunda. Those four paintings were chosen based on the availability of information in the knowledge graph. Multiple stages make database calls for keeping track of conversation coverage of known objects, and deciding when to transition to the next step. The call asks for Item, Content, Material, and Genre nodes in relation to a painting or its description property depending on the stage.

Afterward, we implemented the functions that take care of conversation repair. To address the user having issues noticing or defining relevant items in the painting, we made a function that gives answers to queries based on user input and information from the knowledge base. Next, to allow the user to change the painting they are depicting we made a function that changes the active painting and restarts the discussion from the first topic. Finally, to prevent the conversation from deviating from its goal we overwrote a rasa functionality that takes care of unrelated chat messages and make it encourage the user to talk about the artwork.

For the evaluation of the prototype, we conducted a small pilot test consisting of two participants to identify major issues in the system. The participants pretended to be visitors of a virtual gallery. Afterward, an actual round of testing was planned to begin, which we will talk about in the Measurements section, but will not conduct due to time constraints.

Finally, we used all the previous steps to discuss and hypothesize for the implications and future of the project. Based on this discussion a conclusion was be drawn.

5 IMPLEMENTATION

5.1 Overview

The prototype was implemented using rasa pro/plus 3.8.0b1.dev3, OpenAI's gpt-3.5 turbo model and a Neo4j knowledge graph. On the rasa side, the system ended up with the default DIETClassifier [24] to understand and classify user input, while the overall conversation structure was made with the FlowPolicy and LLMCommandGenerator components. Each stage sends some instructions, the relevant conversation history, and information from the knowledge graph to the API to generate a response. The knowledge graph runs locally via the Neo4j Desktop application. Communication between the bot, API, and database works through custom rasa actions in Python. To detect if a known object was mentioned in the conversation, the word embedding model spaCy was used, alongside scipy for distance calculation, by representing words as vectors, calculating the distance between them and accepting only those within a certain threshold in this case we use cosine distance with a threshold of 0,6.

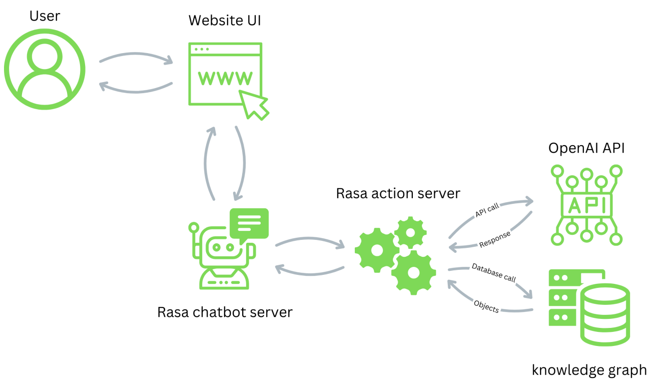

Fig. 3. System Overview

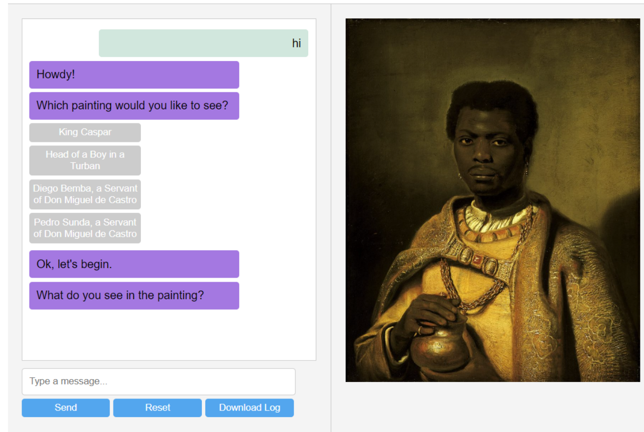

Fig. 4. Chat UI

For the full code refer to Appendix 5.

5.2 Front-end

To conduct user testing in the future, a simple user interface(UI) was made (see Figure 4) using the flask framework [25]. It consists of a chat box, an image container, and three buttons. The image container presents the image based on the current painting. The buttons are Send to send a message alternatively, the user can press Enter, Reset to restart the rasa chatbot and clear the chat box and image, and Log to generate a conversation log in JSON for further analysis.

5.3 Back-end

The rasa server component is the main driver of the conversation, responsible for identifying user intents and determining future actions. Some of the said actions are executed by this component, for example choosing a response from a list, assigning slots, etc. For more complex functionality like API and database calls, we need custom actions written in Python. Those are executed by the rasa action server, which handles the communication between the API and the database.

5.4 Bugs and Errors

The implementation has some issues one of which is the inability to handle multiple messages from the user at a time. Occasionally, the component responsible for deciding the next step in the conversation may flag some input incorrectly and give an inappropriate response. Finally, the function responsible for detecting which known objects have been discussed in the conversation so far, may miss some items or flag them as covered prematurely.

6 MEASUREMENTS

This section describes the testing metrics for evaluation. Note that this paper will not be conducting testing on these measurements. After performing a literature review on papers describing chatbot evaluation metrics [1, 4, 18, 29] and tour guide performance assessment [2, 12] three metrics were extracted: task completion, user satisfaction, and appropriateness of responses.

To begin with, we have task completion to gauge how well the agent manages to improve user opinion forming and observation. User satisfaction refers to the degree to which users enjoyed using the service, with an emphasis on usability and learnability [2, 12] of the artificial guide. This metric, along with the previous one, directly relates to the main research question. Finally, appropriateness will gauge how well the bot responses fit in the conversation concerning VTS principles, relating directly to the first sub-question. This metric will incorporate the chatbots ability to keep the conversation focused on accomplishing the goal and take into account contextual data to detect when the user is giving incorrect information. These aspects will help address the second sub-question.

Task completion could be measured by conducting a controlled experiment with two groups, each with 10-20 testers [28]. The experimental group would describe four artworks while using the prototype, and the control group would use an agent only instructed to make the user describe the painting. The resulting conversation logs will then have to be transcribed and analyzed by experts to see if implementing VTS made any significant improvement in opinion forming. The analysis will consider metrics such as user character count, argument backing by providing evidence, and how many objects in the painting were discussed [11]. To address user satisfaction, a usability test could be conducted with 10-20 participants [8] focused on the UI and chatbot. After said experiment the participants will be interviewed with questionnaire containing 10 questions using the System Usability Scale (SUS). Appropriateness could be evaluated using a combination of analysis of conversation logs after conducting the controlled experiment mentioned above [1] and automated testing methods [4] such as BLEU[23] and ROUGE[15]. For the analysis we could measure the system accuracy [19] defined as the number of responses that are appropriate over the number of all responses. The responses would be classified manually based on context awareness, defined by the ability to responded to inputs regarding an object or person in or not in the painting, dialogue handling, describing competence in responding to the user's requests for assistance or preventing attempts to go off topic, and adherence to the VTS stages outlined in the methods of research section.

7 DISCUSSION

Expressing your opinion in the presence of others can often be a daunting task for people, especially if it is on a subject they are not familiar with. Resulting in a lack of observation of details and a shallow understanding of the art, as these aspects often need time to develop. Although a good guide can make the process much easier, that solution is not a good fit for a virtual gallery application that is always available. Considering the recent advancements in the AI sphere and the insight gained from the literature review, a conversational agent is a good option to explore.

After researching VTS we determined four parts and adapted them into a conversation flow: Observation, Describe story, Describe author technique, and Describe interpretation. Each stage identifies core aspects of an artwork that require careful observation of objects, persons, motives, portrayed feelings, and art techniques from the viewer to encourage visual learning on a higher level. Alongside that design, we used ten requirements to implement a prototype, of which eight were realized. The ones that were not are quality requirements two and seven, due to time constraints and design changes during development. The goal was to foster a discussion with the user while retaining optimal satisfaction. Due to limited time, no large-scale testing was done, however, evaluation metrics alongside testing methods were outlined based on previous research done on similar systems in section 6.

This study aimed to investigate how new technologies like LLMs and personal assistants could be implemented into the art museum space in the role of a tour guide. The novelty in this research comes from trying to incorporate practices used to improve visual learning, namely Visual Thinking Strategies into such systems. Without further testing of the prototype, we can not discern whether or not the implementation manages to accomplish this task. During the pilot test the participants did discuss some of the objects in the painting, however, due to bugs and flaws in the design the experience did not provide insight regarding their knowledge of the painting, or abilities to describe it. Based on this, the design and current functionality of the system we can hypothesize on future implications. In the case that the prototype performs well, it will provide a novel interaction between the user and the exhibits displayed in the virtual gallery. This could increase the satisfaction gained from the experience, and potentially foster an interest in art in general [16] if they did not previously have it. It could also provide the gallery with insight into how their paintings are interpreted by users, which could help in forming the layout of exhibitions by allowing interpretations with similar themes to be organized together for example.

7.1 Challenges

During the design and development of the prototype, there were several problems and difficulties, which will be addressed in this section. Initially, the conversation flow was implemented using rasa stories and rules. However, due to restrictions in functionality, such as lack of branching and looping, flows were employed instead. This approach was not without its issues, key of which was reliably identifying when to trigger the conversation repair functions. For some inputs, such as "I don't know, what do you think" rasa would incorrectly consider a request for help as an attempt to start chatting and tell the user to focus on the painting instead. To remedy this issue ability to answer queries was given by overwriting the in-built function responsible for this behavior. This action helped to reduce the rate at which this issue occurs but did not remove it entirely. Similarly, response generation was planned to be based on triggers in the user's input. From early on, the interaction seemed very artificial, and identifying key triggers was difficult since how art is described will vary from person to person. For those reasons, a pre-trained model was used instead. Another challenge was implementing a known object detection function. At first, a string-matching method was employed. However, that approach was too imprecise and would miss objects if the user did not use words that were close to the examples given. To remedy this word-embedding solution was implemented, which worked better, but could still overlook some words and phrases due to the limitation of descriptions in the knowledge graph. Regarding testing, conducting a proper user study as described in the measurements section would have been ideal, but after taking into account the time needed and discussing it with my supervisors, it was decided that was not feasible. Multiple flaws were identified as a part of the early pilot test. One design flaw was overlooking the possibility of the user asking a question concerning the art piece due to the initial design of the conversation flow being overly concerned with giving curated questions rather than straight answers. Next, one of the testers wanted to change to a different painting in the middle of the discussion, resulting in the prototype starting a new discussion with the wrong image displayed and giving incorrect information. Finally, the same tester attempted to end the discussion early, which was not possible, and the conversational agent continued trying to ask questions. All of the above made conversing with the agent unnatural and machine-like, which was directly against the system's goals and had to be changed. Implementing a solution for each of these problems involved researching rasa patterns that allow for actions to trigger during normal flow execution. Patterns were challenging to work with because of their limitation in linking to other flows after execution, meaning they always return to the flow that the conversation was in before the pattern was triggered.

8 LIMITATIONS AND FUTURE WORK

This section describes aspects the paper failed to cover during research. The following subsections address prototype shortcomings, design flaws, improvements, and future work.

8.1 User testing

Due to insufficient time to conduct a proper user study, there was no large-scale testing of the current prototype. There is a plan in the measurements section describing a potential testing strategy. The next step for this project would be to carry out a study on the impact of the prototype on user opinion forming, in addition to conducting an experiment for user satisfaction.

8.2 Linear design

The design describes a five-stage linear system with the same topics being discussed in the same order for each painting. This pattern is not ideal because the user may begin to see a pattern and become disinterested in the conversation or worse, wish to discuss an aspect of the painting that falls under a different stage. Therefore, in the future, a good point of improvement would be to make each stage link to every other and allow the user to move between them whenever they want.

8.3 LLM intent classification

Initially, we attempted to use the experimental rasa component LLMIntentClassifier for the intent classification and entity gathering. Unfortunately, the results were worse than using the base rasa DIETClassifier, so we chose to use it instead. In the future, utilizing this element properly could introduce an improvement in accuracy and robustness.

8.4 Lack of painting data

Currently, the knowledge graph contains information mainly about the creation and history of the art pieces and less about what is in them or how they were made. That is an issue since the agent uses that information to prompt the user when they are struggling and to keep them on track in case they decide to talk about another painting or describe things that are not in the current one. Additionally, the current means to detect known objects in the conversation does support some variation but can still overlook certain words and phrases. A good subsequent goal would be to provide more data to the knowledge graph as it allows the model to consider more entities when constructing guidance prompts and makes recognition of known objects more reliable.

8.5 AI model

As of writing, the system uses the gpt-3.5-turbo model by OpenAI that already has an improved version at the time of writing. Hence, for future iterations of the prototype or further research, better performance could be achieved by using a more advanced or specialized AI model.

8.6 Integration

Owing to limited time and not being within the scope of the research question, the prototype was not integrated into a virtual gallery exhibition. Communication is entirely text-based, which somewhat takes away from the verbal nature of VTS discussions. A logical next step would be to incorporate the system into an actual art environment and develop means through which it can communicate (voice recognition, text-to-speech).

9 CONCLUSION

In conclusion, this research paper has attempted to utilize novel technologies such as LLMs and conversational agents in combination with Visual Thinking Strategies to develop a prototype chat system. While there is an absence of proper testing, we proposed a clear plan and metrics for evaluation. In addition, we outlined the limitations and improvements for the future of the project. The most important part being executing the testing plan to evaluate the effectiveness and potential impact of the solution.

10 ACKNOWLEDGEMENTS

This paper would not have been possible without the guidance of my supervisors Shenghui Wang and Delaram Javdani Rikhtehgar. Their help played a crucial part in the formulation and execution of this research.

11 APPENDIXES

11.1 Appendix 1: Requirements

Technical requirements:

(1) The chatbot must be built in RASA.
(2) The chatbot must utilize an LLM in either NLU (Natural Language Understanding) or NLG (Natural Language Generation) to improve the experience.
(3) The chatbot must store and extract data from/into a Neo4j knowledge graph.

Quality requirements:

(1) The chatbot must foster interactive discussions about artworks by asking open-ended questions to encourage observation and interpretation, following VTS principles.
(2) The chatbot must collect user interaction data to gain insights on user preferences.
(3) The chatbot must give short, clear and concise prompts to the user.
(4) The chatbot should be aware and intelligent enough to avoid feeling like a questionnaire.
(5) The chatbot should adapt its guidance strategy based on user preferences and past interactions.
(6) If the user is reluctant or having difficulties answering the chatbot must assist them without taking over the conversation.
(7) The chatbot should provide educational content about the artworks, artists, and history, by presenting context and background information while the user is talking about an art piece.

11.2 Appendix 2: Observation/Describe story prompt

['role': 'system', 'content': 'You are a tour guide trained in Visual Thinking Strategies (VTS). You are chatting with a single user who is interested in learning more about the artwork King Caspar.Do not describe the artwork, ask open-ended questions to guide the user through it.If it is the beginning of the conversation, ask "What do you see?". Comment on and commend their observations. You can give only one description and one question per response. Keep responses short if possible. If the user wants to talk about a specific item ask a question about it. Ensure all of your responses transition smoothly with the latest message in the conversation. Always finish with a question. If the user tries to talk about another painting , gently guide them back to King Caspar. You want to do this in 4 stages. The current stage is observation. Vary the beginnings of your responses. Do not say the stage. Here is a list of items with descriptions of what is in the painting: Object: expensive clothes Description: No description available Object: Caspar, Balthaser Description: an african, one of the three magi who came to worship the Christ child Object: jar of incense Description: gift from Heerschop Object: man's face Description: looking proudly and self-confidently If the user mentions an item without describing it, prompt them to describe it. Do not use the description of items when forming a question about them. Do not talk or make up items that are not in the list. If the user begins describing objects not in the above list prompt them with one of them.', 'role': 'assistant', 'content': "Let's get started.", 'role': 'assistant', 'content': 'What do you see in the painting?', 'role': 'user', 'content': 'I see a man']

11.3 Appendix 3: Example conversation

"user-message": "hello"

"bot-message": "Hello!"

"bot-message": "Which painting would you like to see?" "King Caspar" "Head of a Boy in a Turban" "Diego Bemba, a Servant of Don Miguel de Castro" "Pedro Sunda, a Servant of Don Miguel de Castro"

"user-message": "King Caspar"

start with observation

"bot-message": "Ok, Let's start."

"bot-message": "What do you see?"

"user-message": "i see a man"

"bot-message": "Can you describe the man in the painting?"

"user-message": "well he is wearing some pretty expensive looking clothes"

"bot-message": "What do you think the man's clothes say about him?"

"user-message": "that he is wealthy"

"bot-message": "What else stands out to you about the man in the painting?"

"user-message": "he is holding a golden jar"

transition to describe story

"bot-message": "What do you think the man might do with the golden jar?"

"user-message": "i think he is trying to show the jar is important"

"bot-message": "That's an interesting observation! Why do you think the man is trying to show the importance of the golden jar in the painting?"

"user-message": "maybe it has some sort significance to the man in the painting"

transition to author technique

"bot-message": "That's a great interpretation! What do you think the artist's technique is in highlighting the importance of the golden jar in this portrait?"

"user-message": "well it is in the middle and held by the person portrayed in it"

"bot-message": "That's a keen observation! The artist's technique of placing the golden jar in the center and having the man hold it does draw attention to its significance. What emotions or thoughts does this technique evoke in you?"

"user-message": "i dont know what is your opinion"

"bot-message": "I think the artist's technique of placing the golden jar in the center and having the man hold it effectively highlights its importance in the painting. It adds to the sense of wealth and significance surrounding the man."

"bot-message": "How do you think the artist's technique of emphasizing the golden jar influences your perception of the man in the portrait?"

"user-message": "well he looks confident and dignified to me but in general not because of the jar"

"bot-message": "That's an interesting perspective! How do you think the man's confidence and dignity are conveyed in the painting, aside from his interaction with the golden jar?"

transition to describe interpretation

"user-message": "well his expression looks calm and collected which gives me a sense of confidence from him"

"bot-message": "That's a perceptive observation! How does the man's calm and collected expression add to your understanding of his character in the painting?"

"user-message": "i think it is very fitting for a king"

"bot-message": "Absolutely! The calm and collected expression does convey a sense of regal bearing and authority, befitting a king. How does the portrayal of the man in the painting as a king influence your overall interpretation of the artwork?"

"user-message": "I cannot think of any other interpretation"

transition to summary

"bot-message": "That's completely fine! The portrayal of the man in the painting as a king exudes power, wealth, and importance, which are central themes in the artwork."

"bot-message": "Ok let's look at some other paintings."

"bot-message": "Which painting would you like to see?" "King Caspar", "Head of a Boy in a Turban", "Diego Bemba, a Servant of Don Miguel de Castro", "Pedro Sunda, a Servant of Don Miguel de Castro"

11.4 Appendix 4: Tools

During the preparation of this work the author(s) used Grammarly free version, Elicit and ChatGPT in order to edit and improve the flow and grammar of the paper; find research papers; get ideas and coding examples. After using this tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

11.5 Appendix 5: Github

https://github.com/DanielUniTwente/ResearchProject

REFERENCES

[1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. 2020. Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *Journal of Medical Internet Research* 22, 6 (June 2020), e18301. https://doi.org/10.2196/18301

[2] Željko Anđelković, Sanja Kovačić, Marija Bratić, Miroslav D. Vujičić, Uglješa Stankov, Vanja Pavluković, Aleksandra Dragin, Tatjana Pivac, Anđelija Ivkov Džigurski, Ljubica Ivanović Bibić, Zrinka Zadel, and Smiljana Đukičin Vučković. 2022. Museum Tour Guide Performance: A Visitor Perspective. *Sustainability* 14, 16 (Aug. 2022), 10269. https://doi.org/10.3390/su141610269

[3] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. 2017. Rasa: Open Source Language Understanding and Dialogue Management. arXiv:1712.05181 [cs.CL] https://arxiv.org/abs/1712.05181

[4] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ACM, Virtual Event Netherlands, 280–286. https://doi.org/10.1145/3395035.3425319

[5] Beth Clark-Gareca and Tom Meyer. 2023. Visual thinking strategies for English learners: Learning language through the power of art. *TESOL Journal* 14, 2 (2023), e698. https://doi.org/10.1002/tesj.698 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/tesj.698.

[6] I. Dergaa, H.B. Saad, A. El Omri, J.M. Glenn, C.C.T. Clark, J.A. Washif, N. Guelmami, O. Hammouda, R.A. Al-Horani, L.F. Reynoso-Sánchez, M. Romdhani, L.L. Paineiras-Domingos, R.L. Vancini, M. Taheri, L.J. Mataruna-Dos-Santos, K. Trabelsi, H. Chtourou, M. Zghibi, Ö. Eken, S. Swed, M.B. Aissa, H.H. Shawki, H.R. El-Seedi, I. Mujika, S. Seiler, P. Zmijewski, D.B. Pyne, B. Knechtle, I.M. Asif, J.A. Drezner, Ø. Sandbakk, and K. Chamari. 2024. Using artificial intelligence for exercise prescription in personalised health promotion: A critical evaluation of OpenAI's GPT-4 model. *Biology of Sport* 42, 2 (2024), 221–241. https://doi.org/10.5114/biolsport.2024.133661

[7] Elicit. 2023. *Elicit: The AI Research Assistant*. https://elicit.com

[8] Laura Faulkner. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers* 35, 3 (Aug. 2003), 379–383. https://doi.org/10.3758/BF03195514

[9] V. Ferrara, D. Shaholli, A. Iovino, S. Cavallino, M.A. Colizzi, C. Della Rocca, and G. La Torre. 2022. Visual Thinking Strategies as a Tool for Reducing Burnout and Improving Skills in Healthcare Workers: Results of a Randomized Controlled Study. *Journal of Clinical Medicine* 11, 24 (2022). https://doi.org/10.3390/jcm11247501

[10] V. Gupta, A. Sood, and T. Singh. 2022. Disease Detection Using RASA Chatbot. 94–100. https://doi.org/10.1109/MECON53876.2022.9752338

[11] Dabney Hailey, Alexa Miller, and Philip Yenawine. 2015. Understanding Visual Literacy: The Visual Thinking Strategies Approach. In *Essentials of Teaching and Integrating Visual and Media Literacy: Visualizing Learning*, Danilo M. Baylen and Adriana D'Alba (Eds.). Springer International Publishing, Cham, 49–73. https://doi.org/10.1007/978-3-319-05837-5_3

[12] Jinsoo Hwang and JungHoon (Jay) Lee. 2019. Relationships among Senior Tourists' Perceptions of Tour Guides' Professional Competencies, Rapport, Satisfaction with the Guide Service, Tour Satisfaction, and Word of Mouth. *Journal of Travel Research* 58, 8 (Nov. 2019), 1331–1346. https://doi.org/10.1177/0047287518803199 Publisher: SAGE Publications Inc.

[13] C. Ishiguro, H. Takagishi, Y. Sato, A.W. Seow, A. Takahashi, Y. Abe, T. Hayashi, H. Kakizaki, K. Uno, H. Okada, and E. Kato. 2021. Effect of dialogical appreciation based on visual thinking strategies on art-viewing strategies. *Psychology of Aesthetics, Creativity, and the Arts* 15, 1 (2021), 51–59. https://doi.org/10.1037/aca0000258

[14] A. Jiao. 2020. An Intelligent Chatbot System Based on Entity Extraction Using RASA NLU and Neural Network, Vol. 1487. https://doi.org/10.1088/1742-6596/1487/1/012014 ISSN: 1742-6588 Issue: 1.

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[16] Alex Lishinski and Joshua Rosenberg. 2021. All the Pieces Matter: The Relationship of Momentary Self-efficacy and Affective Experiences with CS1 Achievement and Interest in Computing. *Proceedings of the 17th ACM Conference on International Computing Education Research* (2021). https://api.semanticscholar.org/CorpusID:237154831

[17] Santosh Maher, Suvarnsing G. Bhable, Ashish R. Lahase, and Sunil S. Nimbhore. 2022. AI and Deep Learning-driven Chatbots: A Comprehensive Analysis and Application Trends. *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)* (2022), 994–998. https://api.semanticscholar.org/CorpusID:249475706

[18] Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A Survey on Evaluation Methods for Chatbots. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*. ACM, Aizu-Wakamatsu Japan, 111–119. https://doi.org/10.1145/3323771.3323824

[19] Jerome McClendon, Naja Mack, and Larry F. Hodges. 2014. The Use of Paraphrase Identification in the Retrieval of Appropriate Responses for Script Based Conversational Agents. In *The Florida AI Research Society*. https://api.semanticscholar.org/CorpusID:16376786

[20] Lucas Mearian. 2024. *What are LLMS, and how are they used in Generative AI?* https://www.computerworld.com/article/1627101/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html Accessed: 2024-06-28.

[21] Neo4j. 2007. *Neo4j: The Graph Database Platform*. https://neo4j.com/graph-database/ Accessed: 2024-06-29.

[22] OpenAI. 2015. . https://www.openai.com/ Accessed: 2024-06-29.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics, Philadelphia, Pennsylvania, 311. https://doi.org/10.3115/1073083.1073135

[24] Rasa. 2024. Rasa DIETClassifier. https://rasa.com/docs/rasa/reference/rasa/nlu/classifiers/diet_classifier/

[25] Armin Ronacher. 2010. Flask: Web Development, One Drop at a Time. https://flask.palletsprojects.com/en/3.0.x/ Accessed: 2024-06-29.

[26] Akilesh S, Sheik Abdullah A, Abinaya R, Dhanushkodi S, and Rajeev Sekar. 2023. A Novel AI-based chatbot Application for Personalized Medical Diagnosis and review using Large Language Models. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*. 1–5. https://doi.org/10.1109/RMKMATE59243.2023.10368616

[27] Stefan Schaffer, Aaron Ruß, Mino Lee Sasse, Louise Schubotz, and Oliver Gustke. 2022. Questions and Answers: Important Steps to Let AI Chatbots Answer Questions in the Museum. In *ArtsIT, Interactivity and Game Creation*, Matthias Wölfel, Johannes Bernhardt, and Sonja Thiel (Eds.). Springer International Publishing, Cham, 346–358. https://doi.org/10.1007/978-3-030-95531-1_24

[28] School of IT, Mapua University, Manila, Philippines, Mary Jane Samonte, Gabriel Edrick O. Acuña, Luis Antonio Z. Alvarez, and Jeffrey M. Miraflores. 2023. A Personality-Based Virtual Tutor for Adaptive Online Learning System. *International Journal of Information and Education Technology* 13, 6 (2023), 899–905. https://doi.org/10.18178/ijiet.2023.13.6.1885

[29] Bayan Abu Shawar and Eric Atwell. 2007. Different measurements metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap Academic and Industrial Research in Dialog Technologies - NAACL-HLT '07*. Association for Computational Linguistics, Rochester, New York, 89–96. https://doi.org/10.3115/1556328.1556341

[30] S. Tackett, L. Eller, S. Scharff, K.S. Balhara, K.M. Stouffer, M. Suchanek, S.L. Clever, P. Yenawine, S. Wolffe, and M.S. Chisolm. 2023. Transformative experiences at art museums to support flourishing in medicine. *Medical Education Online* 28, 1 (2023). https://doi.org/10.1080/10872981.2023.2202914

[31] Angeliki Tzouganatou. 2018. Can Heritage Bots Thrive? Toward Future Engagement in Cultural Heritage. *Advances in Archaeological Practice* 6 (2018), 377 – 383. https://api.semanticscholar.org/CorpusID:59047745

[32] Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. 2023. The "Conversation" about Loss: Understanding How Chatbot Technology was Used in Supporting People in Grief.. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3544548.3581154

[33] Philip Yenawine. 2013. *Visual Thinking Strategies: Using Art to Deepen Learning Across School Disciplines*. Harvard Education Press, Cambridge, MA.