Design, Generation and Evaluation of a Synthetic Dialogue Dataset for Contextually Aware Chatbots in Art Museums

Inass Rachidi and Anas Ezzakri and Jaime Bellver Soler and Luis Fernando D'Haro
Speech Technology and Machine Learning Group (THAU) - ETSI de Telecomunicación
Av. Complutense 30, 28040, Madrid, Spain - Universidad Politécnica de Madrid
{inass.rachidi, anas.ezzakri}@alumnos.upm.es
{jaime.bellver, luisfernando.dharo}@upm.es
Corresponding author: luisfernando.dharo@upm.es

Abstract

This paper describes the design, synthetic generation and automatic evaluation of ArtGenEval-GPT++, an enhanced dataset designed for training and fine-tuning conversational agents with artificial awareness capabilities targeting the art domain. The dataset build upon the previously released ArtGenEval-GPT, but extended by us to allow more personalization characteristics (including for instance, gender, ethnicity, age, knowledge) and addressing limitations such as low-quality dialogues, and hallucination.

The dataset is generated using state-of-the-art large language models (LLMs), and consists of approximately 12,500 dyadic multi-turn dialogues across diverse museum scenarios, including varied visitor profiles, emotional states, interruptions, and chatbot behaviors. Comprehensive evaluations using objective metrics demonstrate its quality and contextual coherence. Additionally, we explore some ethical implications and limitations of the dataset, such as biases and hallucinations, and outline future directions for enhancing their utility.

These contributions advance the development of personalized, context-aware conversational agents capable of adapting to complex real-world settings, such as museums, while increasing visitor engagement and satisfaction.

1 Introduction

Recent progresses in conversational AI have been achieved thanks to Large Language Models (LLMs), which excel in generating high-quality responses following carefully designed prompt instructions (Sahoo et al., 2024). Then, these same LLMs can be improved further by fine-tuning them on diverse human and synthetic instruction-based datasets and enriched with techniques like reinforcement learning from human feedback (RLHF), showcasing remarkable in-context learning capabilities (OLMo et al., 2024; Abdin et al., 2024; Hurst et al., 2024). This has led to breakthroughs in natural language understanding, dialogue coherence, and adaptability across various domains.

Despite these advancements, there remain significant challenges in creating conversational agents that are not only knowledgeable but also socially competent and context-aware. Specifically, we are interested in the incorporation of artificial consciousness features, situational awareness, or user profiling into conversational systems which are relevant characteristics for improving user experience. Addressing these gaps can enhance user engagement, improve adaptability to dynamic interactions, and enable the development of more intuitive and meaningful dialogues (Della Santina et al., 2024; Graziano, 2022).

In this paper we focus on generating and enhancing the quality and scope of datasets for training such systems. Building on prior work, particularly the ArtGenEval-GPT dataset released by (Gil-Martín et al., 2024)[1] and the methodology described in (Luna-Jiménez et al., 2024), we introduce ArtGenEval-GPT++. This new dataset is specifically designed for training conversational agents in a museum context, where the chatbot acts as a tour guide, tutor, or art expert that adapts to the knowledge level of the visitor, age, ethnicity and even physical appeareance of the visitor. The dataset also includes external situations or interruptions that could happen in a museum. The updated version also includes improved emotional balance, refined and less hallucination dialogue scenarios, and enhanced personalization features such as chatbot tone, multiple visitors and response strategies.

Key contributions of this work include:

- **Dataset Enhancement**: Refining and expanding the dataset to include diverse visitor in-

[1] The dataset is available at (D'Haro Enríquez et al., 2024) and https://huggingface.co/datasets/Astound/ArtGenEvalGPT

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 20–28
May 27–30, 2025. ©2025 Association for Computational Linguistics

teractions, emotional states (i.e., the emotion that the user may have while visiting the museum or aroused when looking into a specific artwork), and contextual situations (e.g., characterstics of the people visiting the museum, unpredicted events that could hypothetically happen while being in the museum, or elements in the paintings that could be used to connect with the visitors).

- **Synthetic Dialogue Generation**: Leveraging the GPT-4-o turbo model to produce 12,500 high-quality dialogues with features tailored to mimic real-world museum interactions.
- **Evaluation Framework**: Implementing automated evaluation metrics, including BLEU, WER, and precision to assess dialogue quality and adherence to prompts.
- **Ethical Considerations**: Addressing challenges such as hallucinations, biases, and limitations in handling sensitive topics .

By focusing on these enhancements, this work contributes to advancing the state-of-the-art in conversational agents, paving the way for more adaptive and socially aware AI systems.). The paper is structured as follows: Section 2 review the seed datasets used in this work. Section 3 details the methodology for dataset design, dialogue generation and automatic evaluation. Section 4 presents the results and examples of the generated dialogues. Finally, Section 5 discusses conclusions and future work.

2 Datasets

2.1 ArtEmis Dataset

The ArtEmis dataset (Achlioptas et al., 2021; Mohamed et al., 2022) is a large-scale resource designed to explore the relationship between visual art, emotions, and natural language explanations. It comprises 455,000 emotional annotations and explanations associated with around 80,000 artworks sourced from the WikiArt website[2]. The dataset contain artworks that span 1,100 artists, 27 distinct art styles (e.g., Baroque, Cubism, Impressionism), and 45 genres (e.g., landscape, portrait, still life). Each artwork includes annotations from at least five human annotators who assigned one of eight predefined emotions or "something-else" as the emotional label, along with textual explanations of their choices.

[2]https://www.wikiart.org/

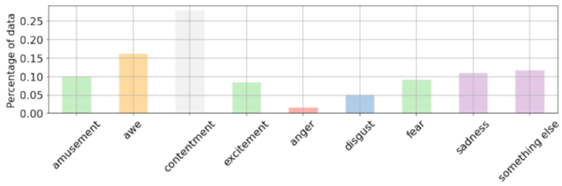

Figure 1: Distribution of Emotions in the ArtEmis Dataset

The dataset's strength lies in its rich emotional and linguistic annotations, which enable a nuanced understanding of the affective and contextual aspects of visual art. However, ArtEmis has limitations in its suitability for training conversational agents, including: Lack of dialogue-specific structures, and limited representation of interactive scenarios or dynamic visitor profiles. Figure 1 shows the distribution of emotions in the original version of the dataset (Achlioptas et al., 2021).

2.2 ArtGenEval-GPT Dataset

According to the authors of ArtGenEval-GPT (Gil-Martín et al., 2024), this dataset was developed as an initial attempt to adapt the principles of ArtEmis for training conversational agents in museum contexts. It introduced a novel synthetic dialogue dataset generated using GPT models, simulating interactions between a human visitor and a chatbot using a flexible platform(Luna-Jiménez et al., 2024). Key features included:

- Dialogue Context: Centered on 800 artworks from ArtEmis.
- Visitor Profiles: Incorporated different age (kid and adult) and knowledge levels (novice, intermedium and expert).
- Chatbot Roles: Simulated chatbots acting as a tour guide, art tutor, or expert.

Despite these innovations, ArtGenEval-GPT faced notable limitations:

- Low Dialogue Quality: Many dialogues exhibited poor coherence due to limitations of the GPT-3.5 model used for generation.
- Hallucinations: Instances of fabricated information about artworks, reducing reliability.

Description	Amount
Total number of generated synthetic dialogues	13,870
Total number of different artworks	799
Total number of different artists	378
Total number of different art styles	26
Distribution of dialogues per emotion	
Emotion	Amount (%)
Amusement	997 (7.2%)
Anger	745 (5.4%)
Awe	943 (6.8%)
Contentment	936 (6.7%)
Disgust	890 (6.4%)
Excitement	885 (6.4%)
Fear	958 (6.9%)
Neutral	6,378 (46.0%)
Sadness	948 (6.8%)
Something else	190 (1.4%)

Figure 2: Key Statistics of ArtGenEval-GPT

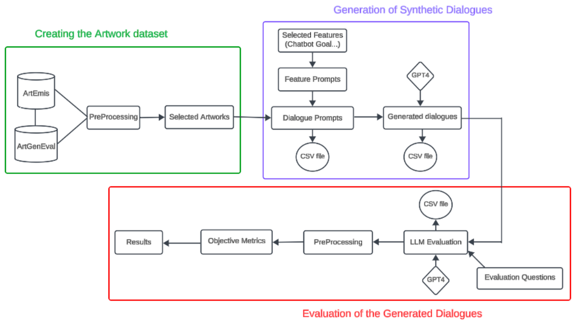

Figure 3: Flow diagram of the methodology used for creating the ArtGenEval-GPT++ dataset

2.3 Limitations and Motivation for Improvements

The challenges observed in ArtGenEval-GPT conducted us to propose the creation of the new ArtGenEval-GPT++. Among the main improvements are:

- Enhancing dialogue coherence by using GPT-4 for generation.
- Perform a better processing of the selected artworks to reduce hallucinations
- Introducing diverse chatbot tones, interruption scenarios, and visitor profiles to simulate real-world museum interactions.

3 Methodology

After inspecting the ArtGenEval-GPT dataset and using it for training our own chatbot, we found the need to improve the diversity of the dataset, and at the same time replace low-quality dialogues (primarily due to limitations of ChatGPT-3.5) and incorporate additional mechanisms to mitigate hallucinations by performing additional preprocessing to the artworks selected for generating the new synthetic dialogues. In addition, with the goal of increasing the capabilities of the chatbot to showcase awareness and extended social competences, we incorporated new ideas and situations.

This section describes the design process. In first place, the new version uses the more powerful GPT-4 model (specifically, GPT4o-mini vs 2024-07-18) and introduce refined preprocessing to improve information accuracy and dialogue consistency. These updates provide new dialogue scenarios with nearly 12,500 new dialogue interactions. 3 summarizes the methodology consisting on three phases: Pre-processing, generation, and evaluation. Note: The indicated CSV files are used to record the output of the different methodology steps for logging purposes.

3.1 pre-processing

The pre-processing step consists of the following steps:

1. **Normalization of artwork information**: In this case, we use Spacy and GPT model to check and normalize that artwork titles and artists were correctly cased and spelled, and remove year of creation from titles.

2. **Filter low quality artworks**: By removing artworks with unknown artists or titles, or artworks whose content is too generic (e.g., landscape, still life, vase).

3. **Keeping high quality emotional artworks**: By removing artworks where the emotion triggered in the annotators was labeled with "something else" or having low inter-annotator agreement.

4. **Verification of artwork knowledge grounding**: In this case, we tasked ChatGPT to return the name of the artist based only on the artwork title (the goal was to detect the uniqueness of the artwork but also to check if the model has knowledge about that artwork). In case, the answer was wrong, we perform a second step by providing also additional information such as year of creation and movement. In case of a correct answer, the artwork is included, if not then it is completely discarded.

5. **Incorporation of additional artworks**: Finally, after removing low quality artworks, but

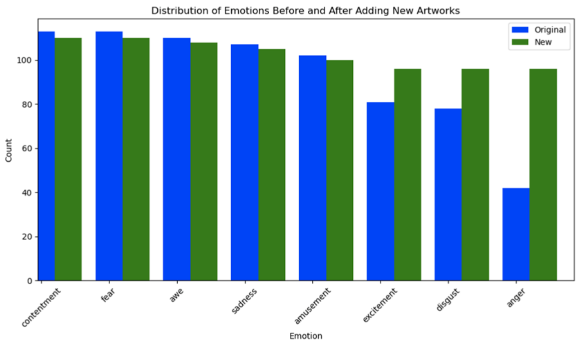

Figure 4: Comparison in the distribution of triggered emotions between the previous and new ArtGenEval-GPT++ dataset.

with the goal of keeping balanced the distribution of triggered emotions, we complemented the list with artworks coming from ArtEmis but also proposed by ChatGPT after repeating the pre-processing steps and performing manual inspection.

Figure 4 shows that after performing the pre-processing steps, the distribution of emotions in the new set of 821 artworks is more balanced that in the previous version, while also improving its quality.

3.2 Generation

The dialogue generation framework customizes interactions by selecting key features such as painting-specific details (title, author, emotion evoked by visitors, date of creation, art style) and dynamic elements like chatbot roles (tutor, tour guide, expert), chatbot behaviour (anthropic or non-anthropic), tone (humorous and playful or professional and formal), and visitor attributes, including gender (male or female), knowledge level (novice, intermediate, expert), age (kid, teenager, adult), emotion (sadness, excitement, etc.), physical/attribute conditions on the visitors (wearing rings, using crutches, using glasses or hat, balding), and ethnicity (African, Asian, Arab, Caucasian, Indian, Latino). It incorporates realistic interruptions and engagement suggestions (e.g., recommending souvenirs, tour options, similar artworks, fire alarms) while supporting group conversations and quick summaries for visitors in a hurry. These features ensure personalized, immersive, and dynamic interactions, addressing diverse visitor needs and enhancing the museum experience.

The dialogue generation process begins by defining the features that vary across dialogues, focusing on visitor preferences, chatbot characteristics, and environmental scenarios. The following steps outline the methodology:

1. **Loading the Dataset:** Artworks are loaded from the filtered and corrected ArtGenEval-GPT v1.0 dataset.

2. **Random Selection of Artworks:** Filtered artworks are randomly selected to ensure diverse dialogues but preserving emotion distribution.

3. **Determining Dialogue Characteristics:** Combinations of chatbot goals, visitor profiles, and engagement suggestions are randomly selected to simulate different museum scenarios. Certain probabilities and rules are implemented to avoid combinations that rarely will happen in real life (e.g., a kid that is expert in art, a group of all people using crutches), while also allowing enough variety of situations and repetitions for the chatbot to learn how to proceed in diverse situations.

4. **Interruption and Engagement:** Random interruptions and engagement suggestions are introduced for realism, with distribution weights applied to scenarios.

5. **Constructing the Prompt:** Selected features are combined into structured prompts, each with a unique dialogue ID, the prompt used, and all feature details.

The prompts are input into GPT-4 to generate unique dialogues tailored to the setup. Responses are saved with token usage and error logs. This approach enables the creation of thousands of dialogues (e.g., 5000 created in a first batch and 7500 in a second batch) with detailed distributions for each feature, ensuring variety and personalization.

3.3 Evaluation

Considering that we generated a total of 12.5k dialogues, we opted for performing an automatic evaluation using GPT-4. This section describes the methodology in terms of prompt design and objective metrics. Considering that we are prompting the GPT-4 model to generate synthetic dialogues that incorporate factual information (e.g., specific information about each artwork), behaviours and characteristics for visitor and chatbot (e.g., tone, demographics, knowledge), as well as length of the dialogues, emotions, situations or even connections between the artwork and visitor's characteristics, we opted for creating specific prompts that could automatically extract that information from the dialogues and then evaluate the quality of the responses. Here, we distinguish between:

- **Exact match attributes** in which exact extration is expected (e.g., for titles, artist name or movement) in which we use Word Error Rate (WER), BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004),
- **Relaxed match attributes** in which information is extracted and compared in terms of Word Error Rate (WER), then a threshold is applied to consider it as valid or not (this allows variabilities when extracting artist names, titles or movements, e.g., Mona Lisa vs Gioconda, or Renaissance vs Renascence), and
- **Behavioral attributes** in which we provide in the evaluation prompt the list of available options plus the N/A option. Then, we check if the retrieved answer is the same as the expected one according to the instructions passed during the generation (e.g., Based on the provided dialogue, determine the visitor gender. **Options:** MALE, FEMALE, N/A.). In this case, the used metric is precision.

The scripts were designed with a systematic approach to ensure consistency and accuracy in processing the retrieved answers. Specifically, we implemented scripts that normalized the obtained responses. This normalization accounted for variations in the answers, as the GPT-4 models did not always provide responses in the expected format, casing, or wording. The scripts also calculated the requested metrics, streamlining the evaluation process.

The implementation process consisted of the following steps:

- **Collaborative Approach:** We begin by preparing a text file that contains questions designed to extract specific information from the dialogues.
- **Focused Inquiry:** Each question is meticulously crafted to pinpoint key details such as the name of the painting and the artist.
- **Direct Communication:** Questions are framed to solicit responses that are concise, direct, and formatted as a numbered list. This format ensures that each answer stands alone, free from unnecessary repetition or additional commentary.

we could categorize our evaluation questions into two main types:

1. **Yes/No Questions:** These questions are designed to elicit direct responses of 'yes' or 'no' from the dialogues. They focus on factual details or specific behaviors exhibited by the chatbot.
2. **Categorical Questions:** These questions prompt responses that fall into predefined categories, such as selecting an emotion or identifying a characteristic attribute. Responses are expected to match one of the provided options accurately. This categorization helps us systematically evaluate how well the generated dialogues adhere to the specified criteria and accurately convey the required information. Additionally, some questions were included specifically to understand the chatbot's decision-making process. In cases where the chatbot does not respond well to the categorical questions, the yes/no questions assist in comprehending how it arrived at its decisions.

4 Results

4.1 Generation

To write our generation prompts, we ensured a clear and distinct structuring by separating factual information about the artwork, the system's role from the visitor's profile, characteristics of groups and general instructions to generate the dialogues (e.g., length, interruptions, connections between artworks and visitors). This organization allows the chatbot to engage effectively with visitors in a museum setting, providing personalized and educational interactions about the artworks. Once the structured prompt dataset was processed and passed to OpenAI GPT-4, it generated the complete dialogue, as partially illustrated in Table 1. The example shows a certain level of emotion (Anger) and lack of time from the user, as well as an attempt of connection considering the age of the user (in this case a kid).

Table 1: Example Dialogue

Visitor: I'm really angry today! Why does this painting look so weird? I have to hurry!
Chatbot: Very quickly, this painting, "The Robbery" by Esaias Van De Velde, shows a dramatic scene. The colors are dark, reflecting the anger and chaos of the moment.
Visitor: But what's with all the people? They look like they're fighting or something!
Chatbot: Yes! The key highlight is the action of the figures. Their movements might remind you of a toy battle, full of energy and tension.
Visitor: I really don't have time for this! I gotta go!
Chatbot: Understood! Just remember, the anger in the painting connects to your feelings today. Have a great day!

4.2 Evaluation

Considering that we have generated a total of approximately 12.5k dialogues, we opted for an automatic evaluation using GPT-4. This section describes the results obtained by applying objective metrics, comparing the actual values of features with the predicted ones.

4.2.1 Categorical Features

For categorical features, we used accuracy as the evaluation metric. Initially, the chatbot behavior emerged as the worst-performing attribute (see Figure 5), which was somewhat expected given that OpenAI is continously refining the development of sentimentally aware chatbots (i.e., avoiding the chatbot to look like a person).

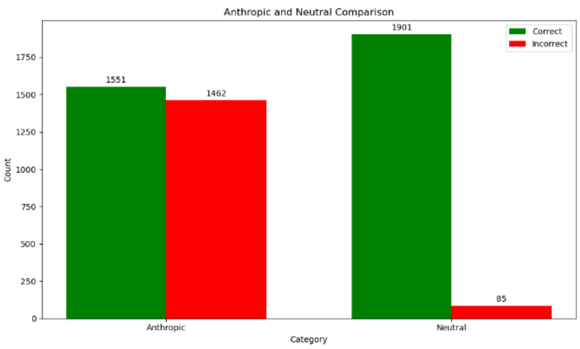

Figure 5: Anthropic and Neutral Comparison.

In some cases, dialogues generated under the Anthropic feature were qualitatively good, but the evaluation metrics failed to capture this accurately due to the subjective nature of attributes like displaying emotions, which are difficult to detect unless explicit emotional cues are present in the dialogue. A summary of our latest results can be seen in Table 2.

Upon closer analysis of the lowest accuracy rates, the most challenging attributes were the visitor's ethnicity and types of interruption.

Table 2: Accuracy of Categorical Attribute Prediction

Attribute	Accuracy
Artist's name	94.7%
Artwork movement	73.3%
Artwork title	86.0%
Artwork year	79.3%
Chatbot role	78.8%
Chatbot tone	81.1%
visitor's knowledge	80.8%
visitor's emotion	86.2%
visitor in a hurry	77.0%
visitor's ethnicity	61.1%
Interruptions type	58.3%

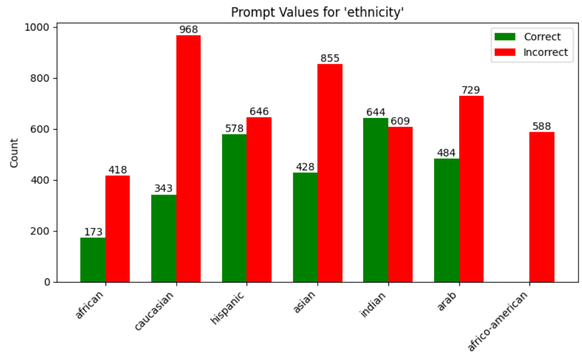

Figure 6: Distribution of the ethnicity detection.

As shown in Figure 6, the most difficult ethnicities to identify were Caucasian and Afro-American. This may be due to strong alingments from GPT models during the generation and for detecting it at evaluation. For interruptions (Figure 7), the most challenging scenarios included detecting that a visitor was not listening to the chatbot (information often omitted in the dialogues) and cases where the visitor was supposed to steal or damage an artwork. Manual inspection revealed that such actions were often attributed to other visitors in the dialogues instead of the intended visitor. When evaluating

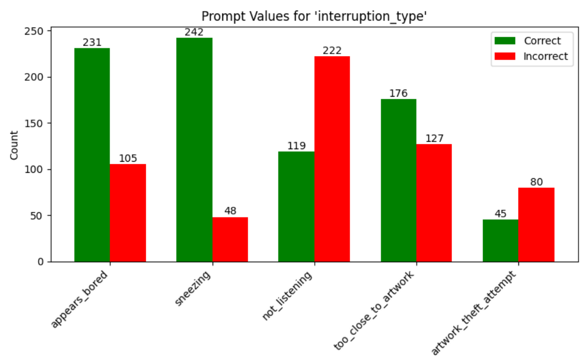

Figure 7: Distribution of classifications for the interruption type.

whether the visitor's gender (male or female) was detected, the accuracies were very low, likely because GPT-4 models are aligned to avoid biases in handling gender information, therefore that information was ommited in most of the dialogues.

4.2.2 Text Attributes

In evaluating text attributes such as painting title, painter name, and movement, we utilized metrics including Word Error Rate (WER), BLEU, and ROUGE to assess the chatbot's accuracy and proficiency. Table 8 organizes results by metric (WER,

Attribute VS Metric	WER ↓	BLEU ↑	ROUGE-1 ↑	ROUGE-L ↑
Painting Movement	0.1112 ± 0.3719	0.6880 ± 0.3884	0.8969 ± 0.2975	0.9032 ± 0.2954
Painting Title	0.0779 ± 0.2523	0.7000 ± 0.3551	0.9324 ± 0.2250	0.9404 ± 0.2189
Painter Name	0.0483 ± 0.2025	0.3086 ± 0.2258	0.9524 ± 0.2003	0.9553 ± 0.1958

Figure 8: Evaluation metrics for textual attributes.

BLEU, ROUGE-1, ROUGE-L) and attribute, highlighting generally strong performance, particularly for painter names. However, the BLEU score of 0.3086 ± 0.2258 was comparatively lower, reflecting the metric's limitations in capturing nuanced similarities in short text segments like names.

To evaluate accuracy, we adopted a WER threshold of 0.2: predictions with WER below this threshold were classified as correct, balancing precision and leniency. This threshold ensured rigorous and practical assessment across all attributes.

Figure 9 illustrates an example of the results obtained for the precision of the painting movement (e.g., Renaissance, Cubism, Baroque, Impressionism) predictions.

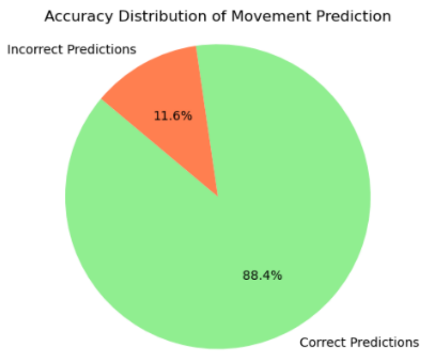

Figure 9: Painting Movement Accuracy.

4.2.3 Single visitor and groups

In the latest version, we included both one-to-one dialogues and group conversations involving the chatbot and two to four visitors. Each visitor had individual characteristics (e.g., gender, age, knowledge) and shared attributes (e.g., ethnicity, emotions). In group dialogues, one or two visitors actively participated to maintain dialogue quality, as having all visitors intervene made it harder to generate coherent exchanges and complicated automatic evaluation.

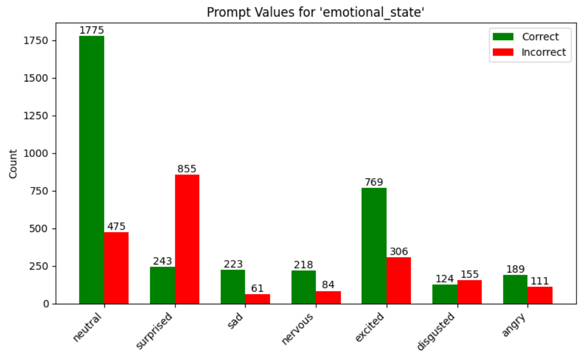

Figure 10: Results on detecting the emotional state of a visitor expressed in the generated dialogues.

Figures 10 and 11 compare emotion detection in one-to-one versus group dialogues, showing it is easier in single-visitor interactions due to GPT-4's difficulty in handling complex prompts and maintaining coherence across multiple visitors. Additional prompt adjustments may improve evaluation of group-based dialogues.

However, it is important to note that, despite the results, correct predictions do not guarantee that all dialogues are free from hallucinations or other artifacts, as these issues cannot be fully detected by the selected methodology.

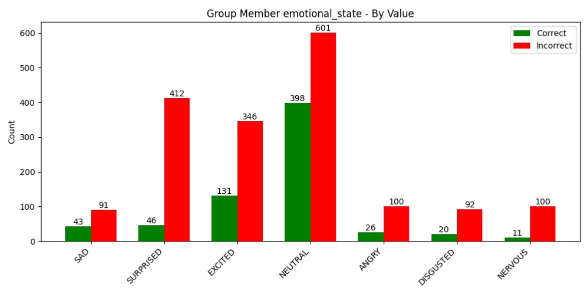

Figure 11: Results on detecting the emotional state of each visitor when considering groups of visitors in the generated dialogues.

5 Conclusions and Future Works

This paper presents a comprehensive approach to designing, generating, and automatically evaluating synthetic dialogue datasets tailored for training aware and socially improved chatbot applications in art museums. Through the refinement of the previous ArtGenEval-GPT dataset, the new version provides a better and extended dataset comprising dialogues over 821 artworks from 384 artists across 26 art styles, ensuring familiarity with GPT models (less hallucinations). By expanding dialogue scenarios to include diverse visitor attributes (including groups, ethnicity, age, physical appearance) and interaction contexts, the new dataset that can be used for training chatbots to engage effectively with varied museum audiences.

Approximately 12,500 dialogues were generated using the latest GPT-4-turbo model, simulating realistic visitor interactions and demonstrating the chatbot's capacity for meaningful engagement. To automatically assess the quality of the generated dialogues, an automated framework was included that relies on objective metrics such as WER, BLEU, and accuracy using LLM-as-Judge, achieving high-quality scores (85.5%). Notable challenges in sentiment prediction and ethnicity detection indicate areas for further improvement, highlighting the complexity of emotional and contextual nuances in dialogue systems. Thus, the dataset provides a valuable resource to advance conversational AI in cultural heritage settings.

Future work will focus on several key areas to enhance unconscious and conscious capabilities. First, expanding the dataset by incorporating a larger number of artworks (including well-known and more emotionally compelling paintings), multiple languages (beyond English, e.g., Spanish, French, German, Italian, Portuguese, Japanese or Chinese), and including the actual image data during the generation to reduce hallucinations and improve contextual understanding. Methodological improvements are also needed by evaluating the tutor scenarios, where the simulated visitors make mistakes in 30% of interactions. This is critical to measuring the efficacy of the chatbot in classroom-like settings for children and teenagers. We are also considering performing human subjective evaluations for evaluating the quality of the dataset and a fine-tuned version of the chatbot. Lastly, based on the Attention Schema Theory (Graziano, 2022), we would like to explore extended reasoning capabilities to model the mind of the visitors, as well as enhancing explainability through tailored prompts. These directions promise to advance the personalization, contextualization, and consciousness of chatbots in our case for cultural heritage contexts.

Acknowledgments

This work is supported by the European Commission through Project ASTOUND[3] (101071191 — HORIZON-EIC-2021-PATHFINDERCHALLENGES-01). In addition, it is supported by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union"; and by INNOVATRAD-CM (PHS-2024/PH-HUM-52) from Comunidad de Madrid.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

Cosimo Della Santina, Carlos Hernandez Corbato, Burak Sisman, Luis A Leiva, Ioannis Arapakis, Michalis Vakalellis, Jean Vanderdonckt, Luis Fernando D'Haro, Guido Manzi, Cristina Becchio, et al. 2024. Awareness in robotics: an early perspective from the viewpoint of the eic pathfinder challenge "awareness inside". In *European Robotics Forum*, pages 108–113. Springer.

Luis Fernando D'Haro Enríquez, Manuel Gil Martín, Cristina Luna Jiménez, Sergio Esteban Romero, Marcos Estecha Garitagoitia, Jaime Bellver Soler, and Fernando Fernández Martínez. 2024. ArtGenEvalGPT. *e-CienciaDatos*.

Manuel Gil-Martín, Cristina Luna-Jiménez, Sergio Esteban-Romero, Marcos Estecha-Garitagoitia, Fernando Fernández-Martínez, and Luis Fernando D'Haro. 2024. A dataset of synthetic art dialogues with chatgpt. *Scientific Data*, 11(1):825.

Michael SA Graziano. 2022. A conceptual framework for consciousness. *Proceedings of the National Academy of Sciences*, 119(18):e2116933119.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

[3]https://www.astound-project.eu/

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Cristina Luna-Jiménez, Manuel Gil-Martín, Luis Fernando D'Haro, Fernando Fernández-Martínez, and Rubén San-Segundo. 2024. Evaluating emotional and subjective responses in synthetic art-related dialogues: A multi-stage framework with large language models. *Expert Systems with Applications*, 255:124524.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. 2022. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21263–21272.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.