CHIM—Chatbot in the Museum

Exploring and Explaining Museum Objects with Speech-Based AI

Oliver Gustke, Stefan Schaffer, Aaron Ruß[1]

CHIM—Chatbot in the Museum is a research project that kicked off with a first brainstorming meeting in January 2020 and ended in June 2022 with the evaluation of the field test data collected with a prototype chatbot application at the Städel Museum, Frankfurt am Main. CHIM will be available as opensource software on GitHub in the second half of 2023. CHIM explores the use of AI-supported, speech-based interactional conversation systems in educational community work in museums. Our aim is to give standard media guides a voice (and a brain). CHIM is a prototype intended to outline a preliminary stage of next-level digital museum guides, which communicate content about museum objects not only 'one way', but instead bi-directionally. CHIM is a step towards future forms of more participatory approaches to communicating with visitors.

The idea for this project was born in a workshop on education in the museum and outreach strategies. One of the participants stated that she normally would not book or use any personal audio guides or media guide systems, because she loves simply browsing among the objects. But then, in front of particular objects, questions sometimes do arise. Questions that are often so specific that she is not able to answer them with a simple, short internet search. That, one might say, 'use story' led to the idea to create a chatbot application for mobile devices that refers to a very specialized database and tries to detect users' intentions in order to find accurate answers: CHIM—Chatbot in the Museum. We began by specifying 15 (in the end, we

1 We would like to thank the Städel Museum, Frankfurt am Main, for their wonderful support, since the CHIM project would not have been possible in the same way without it. CHIM was part of the research initiative KMU-innovativ: Mensch-Technik-Interaktion, which is funded by the Federal Ministry of Education and Research (BMBF) of the Federal Republic of Germany under funding number 16SV8331. CHIM was conducted in cooperation with the DFKI and Linon Medien KG. If someone is interested in joining our chatbot community contact us via chim@linon.de.

conducted testing with 13) artworks at the Städel Museum, Frankfurt am Main, as 'test objects'. We then compiled content and designed a demonstration and test app.

AI as an Assistant in the Educational Field

The connection between the CHIM project and the conference topic is clear: Our research focuses on AI that supports learning in the museum. We thus use AI for natural language understanding (NLU) and natural language processing (NLP) in an interactive conversational system.

Our approach attempts to include the theoretical background provided by John Falk and others in the Museum Visitor Experience Model (Falk 2016, 157), in which the author also points to the important role of identity-related motivations for visits and a Contextual Model of Learning (Falk/Dierking 2018, 135).

Our software prototype is designed for a use case in which visitors are in the museum itself. This means that it has to provide information not only about the exhibits, but also answers to questions like: 'Where are the lockers?' The use case can, however, generally be adapted quite easily to enable visitors to use the chatbot before or after a visit or even outside the museum. For such cases, it is thus necessary to modify the information provided as well.

Key Facts about CHIM—Chatbot in the Museum

CHIM is a software prototype that was developed as part of a research project. Our main aim was thus not to design a fully developed product, but instead to identify best practices as well as challenges related to using NLU/NLP-based AI in the educational field in the museum. Another objective was to establish a community that is also interested in developing speech-based AI-driven technologies that help visitors (or everyone) to better understand objects and themes in our cultural heritage and provide easier access to them. We think approaches like building highly adapted software tools for museums can empower and motivate a larger audience. This means that our prototype has the character of a proof of concept from a theoretical and technical perspective, as well as based on user experience.

CHIM is therefore not yet ready to serve as a product. But considering the impressive speed of development of chatbot applications (like GPT and others), it is plain to see that chatbot modules will play an important role in future education tools in museums. The CHIM approach to providing information about objects, artworks, or even cultural ideas is in many ways 'high level' in terms of data preparation and NLP. A lot of work and resources are, however, required in order to build a proper application. But it is already possible today to provide more 'low level' information

like orientation information or simply the main facts about objects, which are also important for learning in the museum, and offer a more interactive way of doing so by means of a chatbot application.

The CHIM GUI is still relatively raw, with no special branding and *only* a 'standard' chat interface. But we have already implemented a small graphic gimmick: When CHIM computes a certain (poor) probability that the answer the system found is correct, a system message appears, saying: 'I'm not sure if the answer is correct', and we also show an animated GIF, saying the same thing with a meme. We received a lot of only positive feedback from the testers in connection with this graphic gimmick. And, even if this is a little off-topic, since we want to empower everyone working on GUI and interaction design for learning apps in the museum field, our message is: Please, do not forget to make them fun! User interfaces matter, and learning in the museum should be both educational and entertaining.

Technical Implementation

In the first iteration of the project, we developed a sort of speech-based question harvester, with which we compiled over 2000 questions for the 15 'test' artworks on display at the Städel Museum in Frankfurt am Main. We then began by annotating the questions with nine content type categories (in accordance with Barth/Candello/Cavalinet et al. 2020); we subsequently expanded the annotation categories to a total of twelve. The content database consists mainly of audio guide texts and catalogue texts. A small amount of content was, however, specially created for CHIM. We also used the Städel Digital Collection[2] to access basic information about the artworks such as the artist's name or the date of creation. Just as in the case of the questions we compiled, we also annotated the 'answer' content database (not as granularly as the questions) for our twelve content types, so that CHIM would be able to match question and answer intentions.

The NLP stack we used was multitiered and based on several models for detecting intentions and generating answers. We have defined this in greater detail in our previous publications (Zaman/Schaffer/Scheffler 2021a; 2021b; Schaffer/Ruß/Gustke 2023), but we mainly used a RASA[3] model for content type recognition and question-and-answer matching, a document and content type-based matching using models like ELECTRA[4] and BERT[5] (Devlin/Chang/Lee et al. 2019), and cosine-based similarity matching.

2 https://sammlung.staedelmuseum.de/de (all URLs here accessed in June 2023).

3 https://rasa.com/docs/.

4 https://github.com/google-research/electra.

5 https://en.wikipedia.org/wiki/BERT_(language_model).

The first iteration of CHIM was the question harvester developed as a website. As a second iteration of CHIM, we developed an Android app (by using Cordova for app building and MMIR for speech input and output) as well as a Google Chrome extension, which was used solely as a test app. The Android app was installed on (roughly) 15 Android smartphones (Pixel 6, Asus Zenfone 8), along with a customized Android launcher to ensure that testers could see and use solely the CHIM test app. With this test app we conducted a field test at the Städel Museum in Frankfurt am Main from 26 April to 1 May 2022.

CHIM Field Test

The CHIM prototype application provided content for 13 exhibits, ranging from paintings from the fifteenth century to contemporary art objects. Although we initially had 15 objects, two of them were not on display during the test. We wanted the testers to ask at least one question about at least six different objects in one exhibit. For the feedback, we used a standardized questionnaire, called AttrakDiff[6] to obtain users' impression of the GUI/UI. The testers were also able to give open feedback in a text field as well as personally to our interviewers. In addition, we logged user interactions while CHIM was being used in the museum. Overall, we conducted 95 test sessions in which users asked a sufficient number of questions and completed the questionnaire. The testers described themselves mainly as female, with an average age of about 35. We tracked approximately 4,600 user interactions, of which 3,722 were considered in the analysis and 2300 of them questions about the exhibits, which means that every visitor asked on average 25 questions.

The AttrakDiff standard questionnaire provided an overall UX Rating in four dimensions. It has 28 seven-step options for choosing between opposing adjective pairs, for instance, 'confusing—clear'. Each set of adjective items is ordered according to a scale of intensity. This led to a scale value for five different qualities that define the attractiveness of the system:

- Pragmatic Quality (PQ): The ability of a product to satisfy the need for goal attainment by providing useful and usable features. Typical product attributes are: practical, predictable, clear, manageable.
- Hedonic Quality—Stimulation (HQS): The ability of a product to satisfy the need to improve one's knowledge and skills. Typical product attributes are: engaging, creative, original, challenging.

6 https://www.attrakdiff.de.

- Hedonic Quality—Identity (HQI): The ability of a product to communicate messages of self-worth to relevant others. Typical product attributes are: brings me closer to people, expert, connecting, stylish.
- Attractiveness (ATT): Overall positive-negative evaluation of the product: good, attractive, pleasant.

Figure 1: UX rating, CHIM field test.

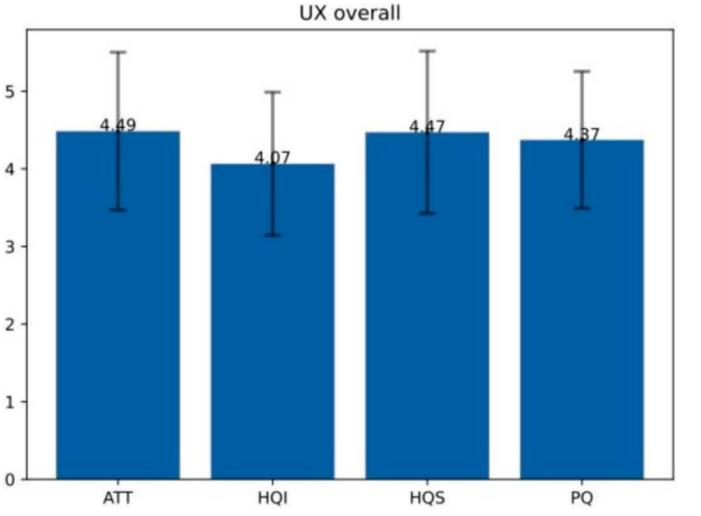


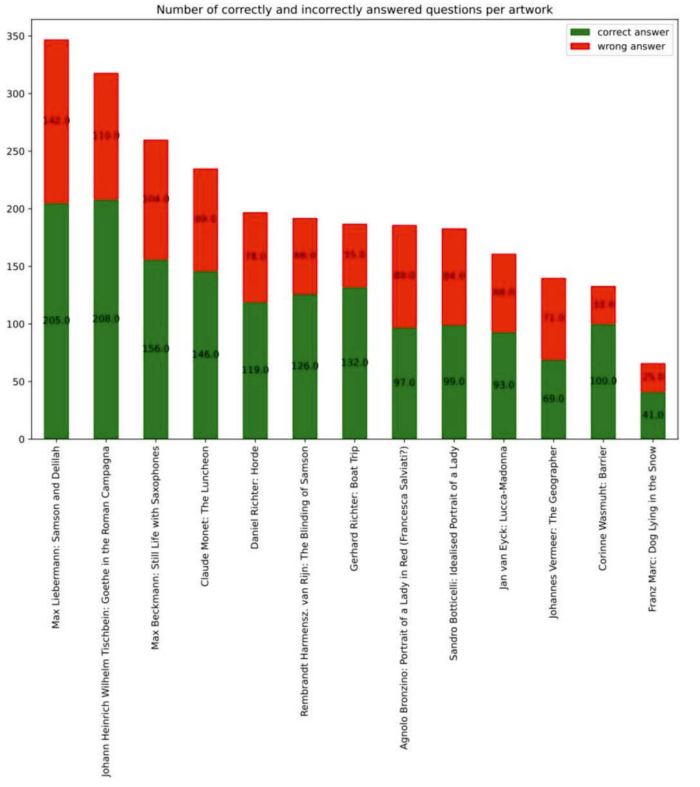
Figure 1 shows that the testers rated the user experience (UX) with the CHIM prototype from 3.63 (HQS) to 4.25 (HQI). For us, this rating is, however, not the most important finding, because CHIM was a test application and not fully developed with respect to the GUI and, furthermore, the sample was not representative. But this rating gave us a sort of general direction and a UX baseline showing that we were on a good path in developing our chatbot-based application, but also that it should be optimized in the future.

By annotating the user interactions, we also analysed the answer quality. Figure 2 shows the answer quality per exhibit—how many right or wrong answers were given—and the total number of questions per exhibit. The number of correct answers is shown in green and incorrect ones in red. In total and in relation to all the questions asked, 63 per cent of all the answers were correct.

Our quantitative data results are a step towards better understanding how museum visitors will accept chatbot-like tools in the educational field and how these applications should be designed. Our studies were explorative and must be confirmed by follow-up studies based on explicit hypotheses (of which we have meanwhile de-

veloped quite a few). From our point of view, there thus is a big chance for further studies, because the system itself and the training data are still available.

Figure 2: Answer quality, CHIM field test.



Lessons Learned and Conclusion

Besides the quantitative data, we also gathered written or oral feedback from the testers. Based on the findings that we obtained over the complete duration of the CHIM project with our focus group and in discussions with our project team, we learned various important lessons:

- Chatbot-powered tools in the educational field in the museum are one way to improve learning in the museum because with a conversational system it is possi-

ble to reach more and/or other audience groups than with common educational approaches (media guides, personal tours)
- Chatbots can help empower visitors because they feel more confident asking questions to a machine than to a real person in a situation with other visitors, or as one tester stated: 'Ich traute mich mehr zu fragen als bei einer normalen Führung' (translation: I dared to ask more than on a normal tour).
- Chatbots can help motivate visitors to learn more about museum objects: quote from the field test: 'Hat mich motiviert genau hinzuschauen, um Fragen formulieren zu können' (translation: It motivated me to take a closer look, to be able to formulate questions).
- Chatbots provide possibilities to archive more visitor participation because the museum is able to provide an automated but personalized dialogue.

To improve chatbot systems like CHIM it is necessary to:

- Learn better how to say 'I do not know' without causing frustration for users. We must improve our dialogue flow and develop strategies so that a museum chatbot behaves in a more humanlike way in communications. Humans know very well how to say, 'I do not know', without causing frustration. This is also a bias problem because users think that a machine, a computer, has to know an answer.
- Tweak the GUI. By developing a GUI that is more personalized we can eventually minimize this bias, because users will consider the chatbot less machinelike.
- Improve the dialogue flow and the interaction design in general.
- Enlarge the content database. Besides a special database, which, from our point of view, is mandatory, we learned, that we should also provide a much broader database (like Wikipedia). What is also needed here is a clear UI based on several more or less trusted databases.
- Use better (synthetic) voices in the text to speech (TTS).

In conclusion: We think that chatbots in the context of museums and arts or cultural heritage should not be omniscient oracles, but instead fun and encouraging companions on tours. They can connect us with other users and motivate us to produce our own content or comments. They are thus a democratic, empowering tool for more participation.

References

Barth, Fabricio/Candello, Heloisa/Cavalin, Paulo et al. (2020). Intentions, Meanings, and Whys. In: María Inés Torres/Stephan Schlögl/Leigh Clark et al. (Eds.). Pro-

ceedings of the 2nd Conference on Conversational User Interfaces. New York, ACM, 1–8. https://doi.org/10.1145/3405755.3406128 (all URLs here accessed in June 2032).

Devlin, Jacob/Chang, Ming-Wei/Lee, Kenton et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019. Available online at https://aclanthology.org/N19-1423.pdf.

Falk, John H. (2016). Identity and the Museum Visitor Experience. London, Routledge. https://doi.org/10.4324/9781315427058.

Falk, John H./Dierking, Lynn D. (2016). The Museum Experience. London, New York/Routledge. https://doi.org/10.4324/9781315417899.

Schaffer, Stefan/Ruß, Aaron/Gustke, Oliver (2023). User Experience of a Conversational User Interface in a Museum. In: Anthony L. Brooks (Ed.). ArtsIT, Interactivity and Game Creation. 11th EAI International Conference, ArtsIT 2022, Faro, Portugal, November 21–22, 2022, Proceedings. Cham CH, Springer International Publishing AG, 215–23. https://doi.org/10.1007/978-3-031-28993-4_16.

Zaman, Md. Mahmud-Uz/Schaffer, Stefan/Scheffler, Tatjana (2021a). Comparing BERT with an intent-based question answering setup for open-ended questions in the museum domain. Konferenz Elektronische Sprachsignalverarbeitung, 247–53. Available online at https://www.essv.de/paper.php?id=1125.

Zaman, Md. Mahmud-Uz/Schaffer, Stefan/Scheffler, Tatjana (2021b). Factoid and Open-Ended Question Answering with BERT in the Museum Domain. In: Proceedings of the Conference on Digital Curation Technologies. Conference on Digital Curation Technologies (QURATOR-2021). CEUR Workshop Proceedings. Available online at https://ceur-ws.org/Vol-2836/qurator2021_paper_2.pdf.