USABILITY AND USER EXPERIENCE OF A CHATBOT FOR STUDENT SUPPORT

Stefan Hillmann, Philine Kowol, Adnan Ahmad, Ruochen Tang, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin
stefan.hillmann@tu-berlin.de

Abstract: This paper describes the usability evaluation of the parts of the CHATU chatbot. The evaluation was conducted with 21 participants. A focus of this paper is the description of the carefully designed evaluation procedure, which aims to avoid textual priming of the participants. The general evaluation procedure can be applied to other speech- or text-based conversational systems, and additional material is provided. The evaluation results show that the usability and user experience of CHATU are positively rated. However, the naturalness and novelty of the interaction are not optimal, and the potential influence of users' experience with LLMs on the evaluation is discussed.

1 Introduction

In the landscape of higher education, the integration of artificial intelligence tools offers significant potential to enhance the educational experience. Among these innovations, the CHATU chatbot, currently in development at the Technische Universität Berlin, is a promising example. This chatbot is designed to aid university students in effectively organizing and managing their academic pursuits.

However, the potential of CHATU hinges on its usability – a critical aspect that determines its effectiveness in real-world scenarios. To ensure that CHATU meets the high standards of user-friendliness and efficiency, it is essential to adhere to established guidelines in usability engineering. In this context, we draw upon the principles of the usability engineering lifecycle [1], which provides a comprehensive approach to developing user-centric interactive systems.

Recognizing the importance of usability, our study emphasizes the need for a thorough evaluation of the usability and user experience of the CHATU chatbot. A key challenge in this evaluation lies in the methodology: employing assessment methods that do not inadvertently prime users with specific words or phrases anticipated by the system and dialogue developers is crucial. This ensures the evaluation accurately reflects the chatbot's performance in a natural, unbiased user interaction scenario.

This paper presents our usability evaluation of our modularized [2] chatbot CHATU. Therefore, CHATU is briefly introduced in Section 2. Our methodology is explained in Section 3 as a potential example for future usability evaluations of text-based interactions with conversational systems. Finally, our evaluation results are presented in Section 4, and both the method and results are discussed in Section 5.

2 CHATU Chatbot

CHATU[1] bundles multiple chatbots (i.e. modules), which allows users to ask questions about general topics related to studying at TU Berlin, information about courses and modules, and

[1] `https://chatu.qu.tu-berlin.de`

 DOI: 10.35096/othr/pub-7076, URN: urn:nbn:de:bvb:898-opus4-70760

Table 1 – Overview on different modules of CHATU, which are represented as different entities in the multi-party chat with the user.

Module	Description
MENTOR	Provides general information about TU Berlin, like the meaning of special terms (Asta, QISPOS), where to find a cafeteria, or what an examination board does.
MOSES	Will answer questions about modules and related courses at TU Berlin. It helps to get course-related information (e.g., time and location) during the lecture time and to plan a semester schedule.
SEKRETARIAT	Answers questions that are otherwise handled by the office of student affairs team. These are the questions about application and enrollment at the TU, academic leave of absence, semester fees, language certificates, and more.
SMALLTALK	Tries to have small talk with the users when they ask questions the other bots can't answer, e.g. about the weather or locations in Berlin.

questions which have been directed to TU Berlin's Office of Student Affairs (cp. Table 1). This chatbot, named CHATU's, is a question-answering chatbot that provides information collected from parts of TU Berlin's website (see [3] for details). In case of multiple potentially fitting answers, CHATU's provides those by buttons from which the user can choose. Furthermore, a small talk component tries to handle out-of-domain requests.

The interaction with the four modules is represented as a multi-party chat. The role model is a chat between multiple people in popular messengers like Signal or WhatsApp. For each message of the user, CHATU selects the responsible module, as introduced in Görzig et al. [2] (Kowol since 2024). Additionally, the user can directly address one of the participating modules using @*name*, e.g., @mentor, as in the last two messages in Figure 1. Each module is represented by its own name and icon in the chat, and the user always knows the message's source. For example, Figure 1 shows the answer of SEKRETARIAT to the last message of the user (here, "I want to study at TU Berlin").

Currently, MOSES is still under development, while the three other modules are already available for students of TU Berlin and other interested persons. All modules are realized with the open-source variant of the RASA[2] framework. While MOSES and SMALLTALK rely on a handcrafted NLU to classify user requests, SEKRETARIAT's NLU, intent structure, and database are automatically generated, as described in Hillmann et al. [3], from information provided on TU Berlin's websites.

3 Methodology

The main goal of our evaluation study was the measurement of usability and user experience as perceived by potential users of CHATU's SEKRETARIAT module (cp. Section 2 and Table 1).

When testing systems based on a reliable NLU-component, e.g., chatbots or spoken dialog systems, it is crucial to avoid priming the participants with task descriptions. This means preventing the leakage of phrases that the dialog system developers expect as typical user utterances [4]. In the following, we describe the procedure of our evaluation, and the related material is available by the OSF[3].

[2] see `https://rasa.com` and `https://rasa.com/docs/rasa`

[3] `https://osf.io/35jhk/?view_only=433180c2150c43aab31cb05cb153b708`

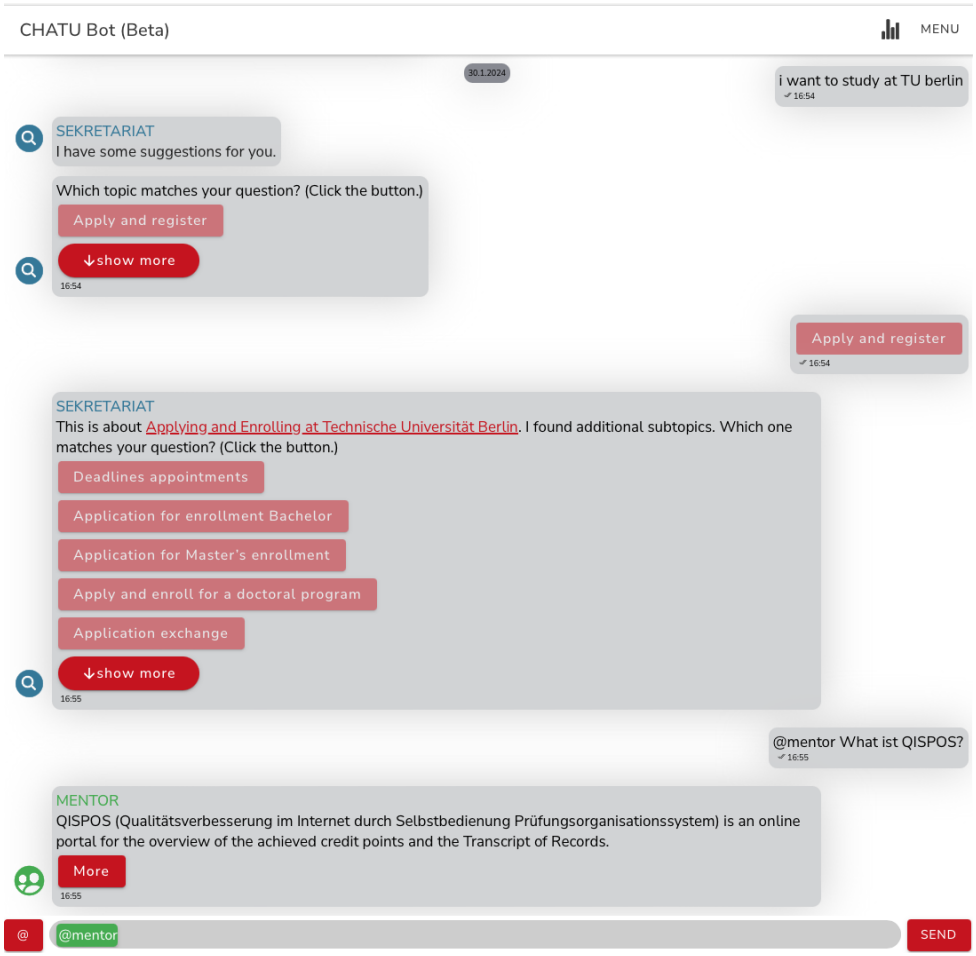


Figure 1 – Screenshot of an interaction with CHATU in English, demonstrating answers based on semantic similarity. The participants could choose their preferred language.

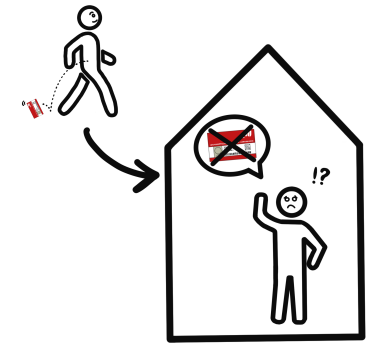
3.1 Target Group and General Procedure

Eligibility for participation required enrollment at a university or universities of applied sciences, proficiency in written German or English, and the ability to attend in-person sessions at our laboratory facilities.

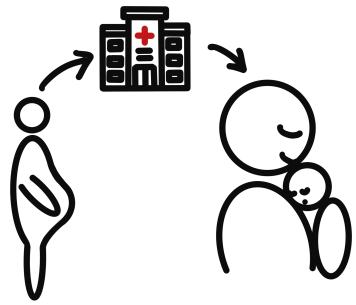
After greeting a participant, we briefly introduced our experiment, provided the formal information sheet, and asked the participant to sign the consent form. Next, a pre-questionnaire, see Section 3.2 for details, about the prior use of chatbots had to be answered by the participant. Then, the experimenter provided the prepared CHATU interface to the participant and started leading the participant through the tasks. Participants typed in their task-related questions or requests for all tasks and sent them to CHATU. The complete procedure, as well as the order of tasks, is depicted in Figure 3.

To avoid priming during the experiment, we used pictures and short conversations to guide participants without articulating the tasks to ensure that the wording was individual. In general, the tasks aim to seek information about the topics addressed in all seven tasks (see Figure 3) by using CHATU. For tasks 1, 4, 5, 6, and 7 in Figure 3, the picture of the related scenario was shown to the participant. Then, they were asked to guess about the situation depicted in the image. The experimenter guided the participants with general confirming or rejecting statements (without using task-related vocabulary) until they stated the expected situation. The experimenter followed written instructions to ensure consistent behavior between different participants.

For Task 2 and Task 3, we guided the participants to the scenario by a conversation. As

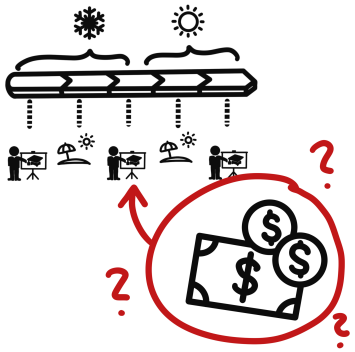


(a) Task 1: Loss of Student ID



(b) Task 4: Maternity Protection

(c) Task 5: Change Study Program



(d) Task 7: Semester Fees

Figure 2 – Images used as stimulus in the tasks 1, 4, 5, and 7. The image for Task 6 (semester break) is similar to the one for Task 7 (semester fees) and is not shown here.

for the other tasks, written instructions described a script the experimenter had to follow in the conversation with the participant. An example conversation of Task 2 (application at TU Berlin) is demonstrated as follows:

> *(EC: Experiment conductor, P: Participant)*
> *EC: What are you doing at TU?*
> *P: Studying media informatics.*
> *EC: What did you have to do to study media informatics here?*
> *P: I enrolled in media informatics.*
> *EC: And before that?*
> *P: I have to apply.*
> *EC: Very good. Let's continue here. Please put yourself in the situation when you did this back then and use the chatbot as if you could have used it at that time to answer your questions.*

The conversation of Task 3 is similar to Task 2, and both tasks were done in an exemplary manner without articulation of the task and without using task-related vocabulary.

After completing all tasks, the participants were requested to fill in three questionnaires (i.e. UEQ, UEQ+, and ITU-T 852, as described below). Moreover, we also solicited individual feedback besides the questionnaires for further improvement of the chatbot. Finally, participants received a compensation of €18 for an overall duration of about 60 to 75 minutes.

3.2 Questionnaires and Metrics

We asked for the usage frequency of classical chatbots and generative AI (GenAI) tools as a proxy for their experience with both. The items formulations are as follows (cp. Table 2 with results for the answer scale):

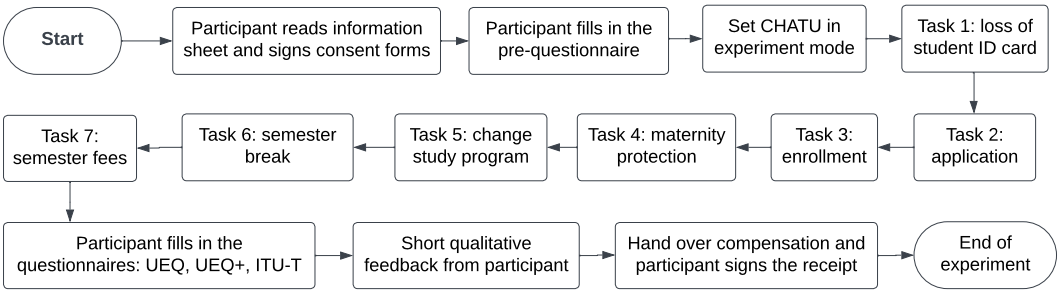


Figure 3 – Workflow of CHATU experiment. Tasks 1, 4, 5, 6, and 7 are completed with pictures as scenario descriptions, while the conductor verbally introduces Tasks 2 and 3 in a conversation.

Table 2 – Transposed table showing usage frequencies of chatbots and generative AI tools (GenAI) in the last 12 months. Frequency abbreviations: Daily (at least once daily), Multi/Week (multiple times a week), Weekly (approximately once a week), Monthly (approximately once a month), Less/M (less often than once a month), Never (not at all).

	Usage Frequency						
Service	**Daily**	**Multi/Week**	**Weekly**	**Monthly**	**Less/Monthly**	**Never**	**∑**
Chatbot	1	1	2	8	6	3	21
GenAI	2	6	5	4	0	4	21

Chatbots: *In the past 12 months, how often have you used a chatbot designed for a specific task? For example, for product information, as a replacement for a support hotline, or for travel information. Hint: Chat-GPT, Bard, and similar so-called generative artificial intelligence are NOT meant here!*

GenAI tools: *In the past 12 months, how often have you used a so-called generative artificial intelligence capable of generating text with which you can communicate in the form of a chat? Hint: Examples of such systems include ChatGPT, Google Bard, Microsoft Bing Chat, Huggingface Chat, or CoPilot.*

For the collection of perceived user experience (UX), we use the extended version of the User Experience Questionnaire (UEQ) [5]. Additionally, we use seven components of the modular UEQ+ questionnaire [6] to get deeper insights into the user experience: Trust, Intuitive Use, Quality of Content, Clarity, Response Quality, Comprehensibility, and Response Behavior.

Furthermore, we used a questionnaire as proposed in the ITU-T Rec. P.852 [7] to get user ratings about the usability of the evaluated chatbot and more specific information on potential usability problems than possible with the UEQ(+) questionnaires.

4 Results

4.1 Participants

In our study, we enrolled 22 participants; however, valid data were collected from 21 of them. Twenty participants reported their age, resulting in a mean of 28.9 years (SD 6.26). When asked for their gender, 14 selected female, 4 male, 2 divers, and 1 preferred not to answer. Furthermore, 12 are students at Technische Universität Berlin, 6 at another institution in Berlin, and 3 at an institution outside of Berlin. Only one participant has used CHATU before the study; the remaining 20 negated that question.

Table 2 shows the participants' general experience with chatbots and generative AI tools. Monthly or more frequent usage has been reported for chatbots by 12 and for GAI tools by 17

Table 3 – Ratings on the UEQ and additional UEQ+ components. The abbreviations stand for: Tr - Trust, Pe - Perspicuity, In - Intuitive Use, De - Dependability, QoC - Quality of Content, Cl - Clarity, Re - Response Quality, Ef - Efficiency, At - Attractiveness, Co - Comprehensibility, St - Stimulation, No - Novelty, and ReB - Response Behavior. SD denotes Standard Deviation.

	Tr	**Pe**	**In**	**De**	**QoC**	**Cl**	**Re**	**Ef**	**At**	**Co**	**St**	**ReB**	**No**
Mean	2.01	1.92	1.87	1.71	1.59	1.57	1.44	1.42	1.31	1.05	0.71	0.40	-0.13
SD	0.74	0.79	0.82	0.66	0.75	1.03	0.88	1.02	0.97	1.21	1.02	1.12	1.49

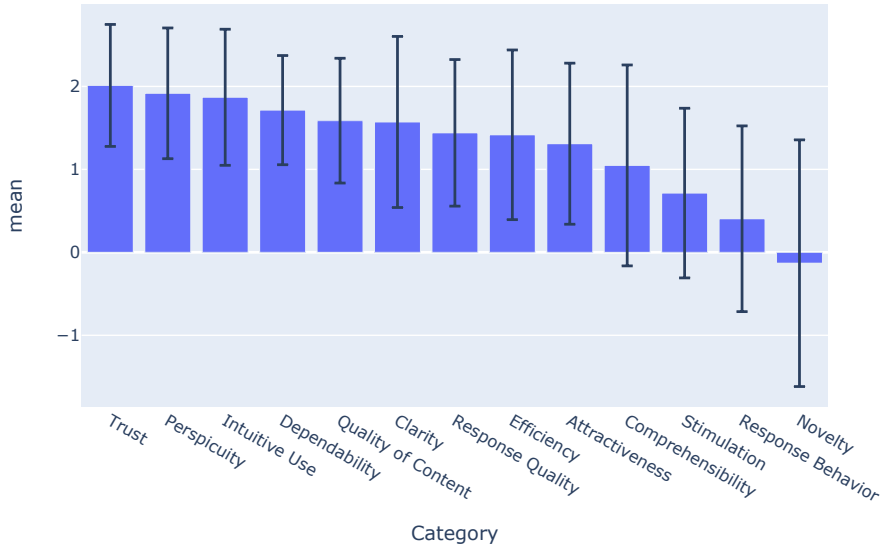


Figure 4 – The bar plot shows mean values and standard deviations of the user ratings with the UEQ questionnaire and additional components from the UEQ+ questionnaire. The categories are ordered by the mean values from higher (left) to lower (right). Higher values correspond with better UX ratings.

participants. Interestingly, GAI tools are used by 8 participants daily or multiple times a week, but chatbots are only used by 2 participants.

4.2 User Experience

Table 3 provides the mean ratings for the different categories of the UEQ questionnaire and the selected UEQ+ components. To ease the interpretation, the same data is shown in Figure 4. While the rating for the pragmatic aspects of CHATU and the interaction are relatively high, the hedonic aspects have – still positive – but lower ratings. Exceptions are Response Behavior and Novelty. For Response Behavior, two out of four items have negative ratings: artificial / natural (mean -0.38, SD 1.69) and boring / entertaining (mean -0.19, SD 1.44). This means the users rate the behavior as artificial and boring – with being close to undecided. Regarding novelty, in the qualitative feedback (cp. last row of Figure 3), several participants compared CHATU with their experience with ChatGPT – which is a high subjective baseline.

4.3 Perceived Usability

Table 4 shows the user ratings collected with a questionnaire according to ITU-T Rec. P.852. Figure 5 visualizes the same data. In all categories, the ratings are positive. The best ratings have been achieved for Provided Information, Communication with the System, and Acceptability. The lowest ratings are achieved for Overall Quality (OQ) and Dialog Flow (DF). In category DF, the item "You perceived the dialogue as natural." received the lowest and only negative rating among all ITU-T items (mean -0.2, SD 0.9). This rating coincides with the low Response

Table 4 – Ratings on the usability questionnaire according to ITU-T Rec. P.852. The table reports mean (and SD) for Overall Quality (OQ), Provided Information (PI), Communication with the System (CS), System Behavior (SB), Dialog Flow (DF), User's Impression of the System (UIS), and Acceptance (AC).

OQ*	PI**	CS**	SB**	DF**	UIS**	AC**
0.62 (0.50)	1.08 (0.55)	0.99 (0.56)	0.78 (0.48)	0.54 (0.58)	1.08 (0.59)	1.13 (0.63)
* Scale:	English: bad (-2), poor (-1), fair (0), good (1), excellent (2) German: unzureichend (-2), schlecht (-1), ordentlich (0), gut (1), ausgezeichnet (2)					
** Scale:	English: strongly disagree (-2), disagree (-1), undecided (0), agree (1), strongly agree (2) German: lehne stark ab (-2), lehne ab (-1), unentschieden (0), stimme zu (1), stimme stark zu (2)					

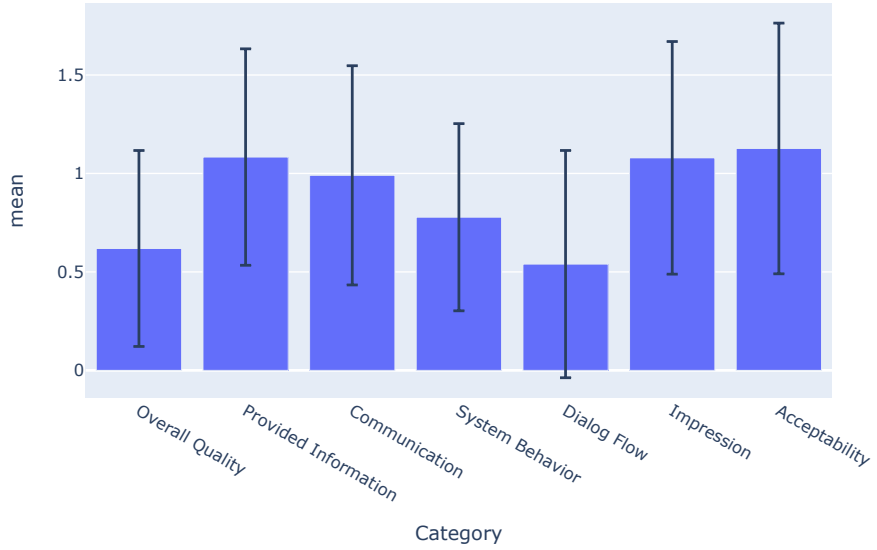


Figure 5 – The bar plot shows mean values and standard deviations of the user ratings with the questionnaire based on the ITU-T Rec. P.852. The complete scale ranges from -2 to 2. Higher values correspond with more positive user ratings. See Table 4 for items of the scales.

Behavior rating in the UEQ+ questionnaire.

Overall Quality is measured by one item ("What is your overall Impression of the interaction?"). The low, still positive rating (mean 0.62, SD 0.5) is in a plausible relation to Dialog Flow, and the as low perceived Novelty (UEQ) is plausible. Again, it is under the impact of the participants' experience with ChatGPT.

5 Discussion and Conclusions

This work only evaluated the perceived usability and user experience, subjectively measured with three questionnaires (UEQ, UEQ+ components, and ITU-T P.852). The study's main goal was to evaluate the user's subjective impression of the interaction with CHATU's SEKRETARIAT module, not the effectiveness (i.e. task success). Therefore, we used tasks known to be solvable with the chatbot and the underlying database. All participants were able to solve all seven tasks without considerable difficulties. The efficiency (effort in means of time and interaction steps) of the interaction will be analyzed in upcoming work.

Besides the generally positive perception of the interaction and the provided information, we see two main outcomes of our study. They address the perceived naturalness (ITU-T P.852)

of the interaction and the perceived Novelty (UEQ), both being low.

On average, the users slightly disagree with the assumption that the interaction with the chatbot is natural. In reflection, this is a plausible result, as SEKRETARIAT is a question-answering chatbot designed for natural dialogs. However, the qualitative feedback gave us different hints on improving this aspect. The chatbot responses have no variance regarding the formulations. More variance in this regard could increase the naturalness of the dialog. Furthermore, the chatbot provides a preview and a link to provide the retrieved information and does not formulate a textual answer to the user's question or request. Here, we are in a conflict between legally compliant answers and natural-sounding answers of SEKRETARIAT. Finally, it might be of benefit if the users could send all messages by typing or using the provided buttons. For now, options on buttons can't be directly selected by written messages.

Chatbots themselves are not novel, and nowadays, a chatbot itself will not be perceived as very novel (as measured by the UEQ) by our target group, (potential) students of our university. Also, based on the qualitative feedback, we hypothesize that the below-neutral rating is impacted by the participants' experience with generative AI in 2023 (i.e. the last 12 months). Many of these tools can be used by a chat-like interface – which is one reason for their success – leading to high user expectations of chat-like systems.

6 Acknowledgment

Parts of the presented work and this paper have been funded by the Federal Ministry of Education and Research (Germany) and the Federal State of Berlin under grant no. 16DHBKI088 for the project USOS at Technische Universität Berlin.

References

[1] NIELSEN, J.: *The usability engineering life cycle*. *Computer*, 25(3), pp. 12–22, 1992. doi:10.1109/2.121503.

[2] GÖRZIG, P., J. NEHRING, S. HILLMANN, and S. MÖLLER: *A Comparison of Module Selection Strategies for Modular Dialog Systems*. In *Elektronische Sprachsignalverarbeitung 2023*, Studientexte zur Sprachkommunikation, pp. 40–47. TUDpress, Dresden, 2023.

[3] HILLMANN, S., P. GÖRZIG, and S. MÖLLER: *Automatic Generation of Website-Based Multi-Turn Question-Answering Dialog Systems*. In *Elektronische Sprachsignalverarbeitung 2023*, vol. 105, pp. 48–55. TUDpress, Dresden, 2023.

[4] EHRENBRINK, P. and S. HILLMANN: *Comparing Priming Effects of Visual and Textual Task Representations - Texts can Influence Users' Utterances*. In *Proc. of IWSDS 2017*. Farmington, US, 2017. doi:10.13140/RG.2.2.27376.66568.

[5] SCHREPP, M., A. HINDERKS, and J. THOMASCHEWSKI: *Construction of a Benchmark for the User Experience Questionnaire (UEQ)*. *Int. J. of Interactive Multimedia and AI*, 4(4), p. 40, 2017. doi:10.9781/ijimai.2017.445.

[6] SCHREPP, M. and J. THOMASCHEWSKI: *Design and Validation of a Framework for the Creation of User Experience Questionnaires*. *Int. J. of Interactive Multimedia and AI*, 5(7), p. 88, 2019. doi:10.9781/ijimai.2019.06.006.

[7] ITU-T: *Subjective quality evaluation of text-based chatbots*. Recommendation P.852, International Telecommunication Union Telecommunication Standardization Sector, 2022. URL `https://www.itu.int/rec/T-REC-P.852/`.